University of Canterbury

Doctoral Thesis

Quantifying Risk of Environmental Exposures Using Bayesian Statistics

---

Author:

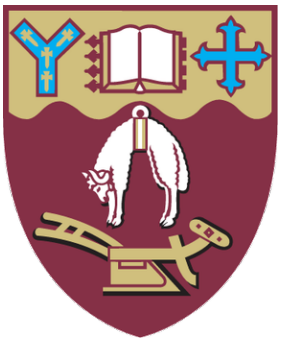
Rodelyn Jaksons

Supervisors:

Dr Elena Moltchanova

Dr Beverley Horn

Dr Elaine Moriarty



A thesis submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

in the

School of Mathematics and Statistics

College of Engineering

2019

"*Whoever loves discipline, loves knowledge, stupid are those who hate correction*"

Proverbs 12:1

UNIVERSITY OF CANTERBURY

Abstract

College of Engineering

School of Mathematics and Statistics

Doctor of Philosophy

Quantifying Risk of Environmental Exposures Using Bayesian Statistics

by Rodelyn JAKSONS

In statistics, the use of observational data is key in understanding what factors are associated with a change in risk. Often, the data also contains a temporal and spatial structure and needs to be accounted for in the modelling. By doing so, one can understand how risk changes over space and time, and to evaluate areas of increased risk. To adequately deal with the complexities of spatio-temporal and observational data, Bayesian hierarchical models can be used. In this thesis, we use Bayesian statistics in the four case studies to quantify risk to environmental exposures and to identify which variables are associated with the greatest change in risk. The applications deal with predictive modelling for water quality management, spatio-temporal analysis of campylobacteriosis disease risk, estimating the extent of underreporting in epidemiological data, and modelling the emergence dynamics of the western corn rootworm for pest management. The models for each application are explained in detail, and the results discussed in depth. Additionally, we also discuss how the methods used in the applications are relevant to other disciplines.

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Deputy Vice-Chancellor's Office
Postgraduate Research Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:
Chapter 7, pages 61-70

Please detail the nature and extent (%) of contribution by the candidate:
Rodelyn did the data cleaning, data analysis, result interpretation and a large part of writing the paper (60%)
The other authors contributed equally (10% each) to writing the paper, advising on subject background and statistical methods.

Certification by Co-authors:
If there is more than one co-author then a single co-author can sign on behalf of all
The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the Doctoral candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Elena Moltchanova* Signature: *Elena Moltchanova* Date: *31/10/2019*

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Deputy Vice-Chancellor's Office
Postgraduate Research Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:
Chapter 8, pages 69-77

Please detail the nature and extent (%) of contribution by the candidate:
Rodelyn was responsible for planning and implementing the data analysis, interpreting the results and writing the paper (70%).
The other authors contributed in equal measure (10% each) to planning of the analysis, consulting on subject background and statistical methods, and the manuscript write up.

Certification by Co-authors:
If there is more than one co-author then a single co-author can sign on behalf of all
The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the Doctoral candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Elena Moltchanova* Signature: *Elena Moltchanova* Date: *31/10/2019*

Acknowledgements

Gratitude must be given to my supervisors Elena Moltchanova, Beverley Horn and Elaine Moriarty. Without your mentoring and patience, the work of this thesis would not have been possible. Thank you all for reminding me to first look after myself, and most importantly, to enjoy the journey. A special thanks must also be given to Elena, who challenged me in my statistical thinking and shaped me as a researcher.

A massive thanks needs to be given to my wonderful family, who throughout the years, had no idea what I was doing, but believed in me regardless. I am thankful for your feigned interest in why probabilities rule our lives. Without you all, the sun would not have shined so bright. I am forever grateful to my parents, Rodrigo and Yolanda, and my parents-in-law, Lieve and Robert, who happily looked after my children so that I could pursue my research.

I am also indebted to my wonderful and loving husband, Peter Jaksons, who supported me throughout this roller-coaster of a journey. Thank you for encouraging me in all that I do, and for always believing in me, it is an incredible feeling to have you by my side. I also wish to acknowledge my two lovely children, Felicity and Wolf, who taught me life is too short to sweat the small stuff.

To the staff and postgraduate students of the School of Mathematics and Statistics, thank you for making my years at UC enjoyable. Thanks for the coffee breaks, biscuits, colouring, crocheting, yoga, lunch-time runs and the welcome comedic relief on a day-to-day basis. I would mention you each by name, but in fear of forgetting at least one person, I shall refrain.

Last but not at all least, credit must be given where it is due. Eternal gratitude must be expressed to our Heavenly Father above. Without my faith in you, all hope and love would have been lost.

In similar words of the great Curtis Jackson: accumulate riches or perish in the attempt. Thus, I will attempt the former until I am the latter.

Contents

List of Figures

List of Tables

List of Abbreviations

AIC	**A**kaike **I**nformation **C**riterion
BN	**B**ayesian **N**etwork
BYM	**B**esag **Y**ork **M**ollie
CAR	**C**onditional **A**uto **R**egressive
CAU	**C**ensus **A**rea **U**nit
CFU	**C**olony **F**orming **U**nit
DIC	**D**eviance **I**nformation **C**riterion
GDD	**G**rowing **D**egree **D**ays
GR	**G**elman **R**ubin
HPD	**H**ighest **P**osterior **D**ensity
MAUP	**M**odifiable **A**real **U**nit **P**roblem
MCMC	**M**arkov **C**hain **M**onte **C**arlo
MH	**M**etropolis **H**astings
MLE	**M**aximum **L**ikelihood **E**stimator
MPN	**M**ost **P**robable **N**umber
SDHB	**S**outhern **D**istrict **H**ealth **B**oard
WCR	**W**estern **C**orn **R**ootworm
WGS84	**W**orld **G**eodetic **S**ystem 19**84**

For Rodrigo and Yolanda, who worked hard so their children could have the opportunities they did not have.

Introduction

The word risk has a negative connotation, as the possible unwelcome consequence of an action or an event. To mitigate risk, we often identify the main risk factors and ensure that exposure to harm is minimised. In this thesis, we analyse spatio-temporal data. Using a range of statistical tools, we identify variables which increase risk, with hope that the insights we provide are useful in reducing negative impacts.

In epidemiology, disease risk is known to rise with increased exposures to different transmission pathways. For example, the incidence of gastroenteritis diseases, such as campylobacteriosis, increases when people are exposed to polluted waters. In horticulture, growers are at the mercy of nature, where climate conditions dictate the yield potential of a growing season, and where they also have to manage the threat of crop diseases and pests. For maize growers in Europe, the western corn rootworm (WCR) beetle is one such pest, and understanding how its emergence dynamics depend on climate is vital for minimising yield loss.

The applications covered in this thesis deal with (I) issues relating to water quality prediction, (II) the spatial analysis of campylobacteriosis incidence, (III) the estimation of disease risk in under-reported data, and (IV) the emergence dynamics and spatial distribution of the western corn rootworm (WCR) beetle. Henceforth the case studies shall be referred to as Study I, Study II, Study III, and Study IV.

Each data set includes a temporal component. In Study I, the aim was to predict the

water quality state for each week of the bathing season. Study II aimed to investigate how campylobacteriosis risk evolved through time, and to make a comparison between urban centres and rural areas. In Study IV, we modelled the emergence dynamics of the WCR beetle over the maize growing season.

Another component which was shared by most data sets is the spatial structure, which was present in the Study II, Study III and Study IV data sets. Therefore, spatial effects were included in these case study models, to describe the observed patterns better and to identify areas of increased risk.

Project Backgrounds

In this chapter, we discuss the case study backgrounds and motivate each analysis.

2.1 Reducing the risk of illness from freshwater swimming

In 2016, when 5500 of the 14,000 residents of Havelock North fell ill with the gastroenteritis campylobacteriosis, authorities were quick to establish that the water supply had been contaminated [121, 89]. Consequently, there was an urgency to stop the further spread of disease and to identify the source of contamination. Researchers found that two water bores, which neighboured nearby paddocks, were contaminated by sheep faeces. This contamination most likely occurred after a period of heavy rainfall, where it inundated the neighbouring paddocks causing water to flow into a nearby pond. The water in the pond then entered the aquifer and flowed to the water bores [89]. This incident highlights the fact that degraded water quality is detrimental to public health as it has the potential to impact the areal population.

Recreational water quality is currently a topical issue in New Zealand, as polluted waters have the potential to cause outbreaks of gastroenteritis and respiratory illnesses [12, 189]. In the summer, popular swimming holes, rivers and beaches usually bring a plethora of families and tourists to cool down in the heat [19]. Historically, there was little concern about the water quality of swimming areas in New Zealand, as they were known to be pristine. However, in recent years, the rise of farming activities and

increased urbanisation have degraded water quality [55, 131, 98, 120, 158, 38]. For some people, a quick dip now resulted in gastroenteritis or respiratory illnesses. The possible negative consequences of the once innocent swim have caused a public outcry. In response, the government updated the national policy statement for freshwater management in 2014, which sees regional councils responsible for maintaining recreational water bodies [116]. They are also in charge of warning users when water quality has degraded.

Water is a mode of transport for many pathogenic microorganisms and indicator bacteria such as *E. coli* are used to signal water quality degradation [48, 142]. In New Zealand, the Microbiological Water Quality Guidelines for Marine and Freshwater Recreational Areas 2003 outlines the acceptable water quality for locations designated for recreational use. The bathing seasons recreational water quality grades; acceptable, alert and action, are assigned based on *E. coli* concentration [115].

Water quality is known to fluctuate rapidly, and due to financial and time constraints, water sampling of the recreational area cannot be undertaken on a daily or more frequent basis. As a result, authorities around the world have seen the need to implement predictive models to fill in the gaps. For example, linear regression models have been used to predict water quality in the USA, United Kingdom, and Hong Kong [183, 60, 59, 36]. Regression trees have been used to predict bathing suitability throughout Scotland, and to predict river quality in Slovenia [175, 46].

Santa Monica beach is located along the Pacific Coast Highway in California, USA. Each year the beach has millions of visitors who are enticed to the water, to cool down from the heat. Therefore, water quality maintenance is paramount, and informing the public when harmful pathogens have been found is vital in reducing outbreaks of gastroenteritis. In 2014, researchers investigated various models that could predict water quality promptly and would perform better than the naive model that was used at

the time [184]. The naive model assumes that the best predictor for today is the water quality grade from yesterday. They compared performance between five statistical models; multiple linear regression, logistic regression, partial least squares regression, artificial neural networks and classification tree and found that these statistical models all performed better than the naive method.

The objective of Study I was to find a model that could predict future *E. coli* counts, or water quality grades based on preceding data in the same season or year. The prediction would be based on past values of *E. coli* counts, accumulated rainfall of a monitored upstream site in the past 48 hours, and river flow. The results of this study provides a basis for a model suitable for real-time prediction for bathing sites across Southland, New Zealand. The proposed model should be able to correctly identify grade action (poor water quality) days, or predict higher levels of *E. coli* concentrations so that the public can be informed of water quality degradation. Such a tool could be used in a local councils website, where it can automatically predict swimming suitability based on meteorological and hydrological data or forecasts.

In Study I, we apply a variety of statistical models, including log-linear regression model, logistic regression model, discriminant analysis, regression trees, random forests and Bayesian networks to predict water quality and discuss their predictive performance.

2.2 Understanding campylobacteriosis risk

Each summer, the New Zealand Food Safety Authority launches a campaign reminding people about the importance of food safety. Cross-contamination of food, such as the mishandling of raw meat, is also known to cause gastroenteritis such as campylobacteriosis. Campylobacteriosis is a type of gastroenteritis caused by the *Campylobacter* bacteria, where *C. jejuni* is the most common strain. Common symptoms are fever, headache and diarrhoea. The disease is acute, with some people experiencing serious sequela to the initial illness.

From the 1980s until the mid-2000s, campylobacteriosis incidence increased in such drastic numbers, that it was described as reaching epidemic proportions [162, 123, 14]. A large study initiated in 2005 indicated that more than 50% of all campylobacteriosis cases were linked to poultry consumption. Consequently, the New Zealand Food Safety Authority and the Poultry industry introduced changes to risk management strategies in 2006 [162]. The changes saw the annual notification rate drop from 358.8 per 100,000 (2002-2006) population to 161.5 (2008) per 100,000 [132, 162, 88]. However, even after the regulatory changes, there remained differences in campylobacteriosis notification rates between urban and rural populations, with rural areas having larger notification rates compared to urban ones [162, 167, 61].

Previous studies have shown that campylobacteriosis risk is different for urban and rural populations, as they can have different pathways to exposure of campylobacter bacteria [97, 102]. For example, people living in rural areas are more likely to be employed in the agriculture sector. Thus, they have a higher exposure to farm animals, which are known to be a major reservoir of *Campylobacter jejuni* [114, 130, 170]. In 2008 the association between New Zealand dairy farming and campylobacteriosis was investigated. The study recruited seven people who had laboratory-confirmed campylobacteriosis, and who lived and worked in dairy farms. The results of the study showed that four of the seven campylobacteriosis cases were likely due to contact with dairy cow faeces [70]. In another dairy farm environment study, *Campylobacter jejuni* was detected in 66% of bovine samples, including 59% of dairy cow and 75% of calf samples [69]. The results of these studies show that rural areas have different disease risk exposures to urban areas. Therefore, the regulatory changes implemented in 2006 may have had different effects in urban and rural populations.

The objective of Study II was to model differences in campylobacteriosis risk between urban and rural areas and to study the spatial distribution of the disease. We were also interested in whether the spatial distribution changed over time, and to determine whether high-risk regions remained the same over the study period.

The data were of notified cases of campylobacteriosis in the Southern District Health Board (SDHB) of New Zealand, for the years 2000 to 2015. It was sourced from EpiSurv, which is a surveillance program that collects information on notifiable diseases from public health services [52]. The reported cases were geo-coded, which provided address accuracy down to census area unit level (CAU). A CAU is a non-administrative boundary and is an aggregation of mesh blocks [171]. Census information on CAUs contained information on population size, as well as an urban or rural classification assigned by Statistics New Zealand [172, 173].

To account for the temporal evolution of the disease, a piecewise linear regression model was used. A separate curve was fitted for urban and rural areas to account for differences in risk. The Besag York and Mollie conditional autoregressive normal (BYM CAR) prior was used to account for the spatial autocorrelation in the distribution of the disease prevalence [23].

2.3 Underreporting of disease risk

To assess the burden of many diseases, researchers often use registry data that monitors new or existing cases of a disease [87, 181, 20, 41, 77]. Based on epidemiological monitoring data, one can study the association between the disease and different risk factors. However, epidemiological monitoring data can underrepresent the actual number of cases for a variety of reasons.

Underrepresentation can happen at random and result in underestimation of overall risk but accurate on-average estimation of association between risk factors and the disease incidence or prevalence. For example, because campylobacteriosis often passes quickly without any treatment, many people do not seek medical advice, and so are missed in the reported cases [58]. However, another reason for not reporting a case may be due to personal finances or living far away from a population centre, where a visit to the doctor is unaffordable and so is foregone [159, 18, 27]. In this context, the mechanism that is responsible for the unknown observations is systematic, and

therefore can introduce bias into the estimates, as it skips characteristics of the missing individuals or groups [157]. It is thus essential to correct for the unreported cases to estimate the true number of people afflicted with a disease. If there is knowledge as to what drives the underreporting, we can include it in the estimation process, thereby adjusting for the unseen cases [177]. For epidemiological studies, knowing the number of people infected with a disease is essential. It enables authorities to understand the overall disease burden, and if the measures for treatment and prevention have been effective [52].

Underreported data is not unique to epidemiology and is known to occur in all disciplines. In criminology, it is known that the many crimes are unreported, with some neighbourhoods and minority groups being underrepresented in the overall figures [136, 79, 178]. In ecology, researchers often want to know the population of a species to see if their efforts in pest eradication or species conservation have been successful. However, as animals are known to move, they may be hard to detect. Because it is impossible to observe the entire habitat at a high enough resolution simultaneously, it is difficult to capture the entirety of the population [108, 154, 153]. Underreported data, may result from the planned sample locations not being visited, as they are challenging to access [129, 152, 42]. Not covering the entire spatial domain is problematic, as the difficult to access areas, can be prime locations for the said species.

In Study III, we use a Bayesian hierarchical framework to model underreported counts, when it is assumed the extent of underreporting is known. We show that the observed cases come from a reparameterised binomial distribution. The binomial parameter, the probability of success, is a product of the disease risk and detection probability. Using simulated scenarios, we show how the model can adequately estimate the effects of different risk factors. The model can also be used to estimate the unobserved number of cases, as well as uncover areas where underreporting is severe. As the model makes use of linear predictors, incorporating spatially autocorrelated errors is straightforward.

2.4 Reducing the risk from western corn rootworm (*Diabrotica virgifera virgifera*)

Like epidemiological monitoring, ecological pest monitoring data collects information on invasive species, such as insects, to quantify the severity of infestation and to gain insight on how far it has spread [91, 35, 127]. The objective of Study IV was to model the western corn rootworm (WCR) beetle emergence dynamics using ecological monitoring data, and to investigate how the dynamics were affected by climatic variables. This study is an extension of the work by [54], which used a zero-inflated Poisson model to assess the severity of WCR beetle infestations in different climate scenarios.

For many growers and farmers, the yield of the crops that they grow determines feast or famine. When crops return high yield, they can sell the fruits of their labour to the market, with the profits enabling them to invest in machinery, or provide their families with financial security. For growers, it is essential to grow crops that consumers demand and that has a high yield potential. However, choosing which crop to grow is not a simple task. Crop rotation may incur additional costs and capital investment. On the other hand, growing the same crop in the same area each year may degrade the quality of the soil, which results in poor yield. In Austria, maize is the second most important crop after wheat. It is a favourite amongst growers, as there is low labour input, and the yield potential remains the same for subsequent years [163, 53].

The western corn rootworm (WCR) beetle originates from the USA and is a major agricultural pest. It is reported that WCR beetle infestation can result in up to 90% maize yield loss, and revenue loss of USD1 billion per annum [182, 57, 39, 73]. In the USA, strategies like crop rotation, in which growers plant corn in alternative years, have proven unsuccessful, so genetically modified varieties of corn are being grown [185, 73].

In Europe, the WCR beetle was unknown until it was first detected near Belgrade (Serbia) international airport in 1992. It was believed to have been accidentally introduced

during the Yugoslav Wars in the early 1990s [96, 113]. It was given quarantine status in the European Union (EU) in 2003, and monitoring schemes were initiated in hopes of eradicating the WCR beetle population [1, 34, 31]. However, eradication was not achieved, and in 2014 the quarantine status was lifted. Growers are now instructed to implement crop rotation and use insecticides to control the WCR beetle population [149, 57, 180, 44, 112].

The WCR beetle is known to be abundant in warm and dry climates, and with the warm nights and sufficient rainfall for maize production, Austria is a prime breeding ground. There is some variability in climate over the country so that one would expect different maize yield potential in different areas, and also varying levels of WCR beetle infestation. The Austrian climate is seasonal. Usually, the coldest month is January, with snow cover in the valleys lasting from December through to March. Often, July is the hottest month and temperatures can exceed 30°C. The lowland regions in the north and east have colder winters and hotter summers with moderate precipitation throughout the year. On the other hand, western areas experience mild winters and warm summers, with higher rainfall. The southeastern regions experience longer and warmer summers [50].

Warm temperatures are essential for plants and insects to mature [200, 25, 7, 6]. After a sufficient number of warm days, emergence cycles for many insects begin to adhere to the following process; a period of no or little growth, followed by a rapid increase, following which growth slows before stopping completely. Mathematically, this process can be expressed by many parametric curves, one of which is the graph of the Gompertz function [72]. The Gompertz curve usually depends on three parameters; the asymptote, a relative starting value, and the growth rate coefficient [186]. The asymptote can be a proxy for the carrying capacity of the species, while the relative starting value indicates the time to first emergence. The growth rate coefficient affects the slope of the growth, where lower values indicate protracted periods of infestation. As emergence dynamics are known to correlate with climate, so there is a need to understand how variables such as temperature will affect population sizes and habitat suitability [107,

180, 9].

In Study IV, we model the emergence dynamics of the established WCR beetle populations through a Gompertz function in a Bayesian hierarchical model. We model the asymptote parameter and growth parameter as functions of known covariates, such as temperature and precipitation, to investigate how climatic variables affect the emergence dynamics. We interpret the asymptote parameter, as proxy to the carrying capacity. The growth rate parameter represents the emergence rate. As the WCR beetle is an established pest species in Austria [54, 40, 47], there is a need to understand its emergence dynamics, so that practices like insecticide spraying are optimally timed to reduce population size [66, 103].

Statistical Models

3.1 The two schools of thought

In statistics, there are currently two prevailing schools of thoughts/philosophies; frequentist (or classical) statistics and Bayesian statistics. The fundamental difference between frequentist and Bayesian statistics is in how the unknown parameter is perceived. In some cases, the parameter may already be a fixed quantity. In other situations, the value of the parameter have been generated from a random process. In Bayesian statistics, all unknown quantities in the model are treated as random variables. Bayesian statistics incorporates prior belief/knowledge about the unknown parameter, thus expressing uncertainty around the parameters value [74]. On the other hand, in frequentist statistics, the unknown parameter is treated as fixed, i.e. deterministic. Inference in a frequentist setting produces a point estimator of a parameter, and uncertainty is expressed through confidence intervals. In contrast, Bayesian inference gives a distribution for the parameter of interest.

In many cases, Frequentist statistics is usually based on the likelihood function, and on the notion that experiments are infinitely repeatable under identical conditions. Thus, the statistics that are obtained from frequentist statistics represent the long term frequency, i.e., the probability of an event. In Bayesian statistics, experiments are considered unique, and prior belief about the occurrence of an event is formed to make probabilistic statements. Therefore in Bayesian statistics, inference is based on the posterior distribution, which is a combination of the likelihood function and prior distribution of the parameter of interest. The prior distribution should be formulated

before the current experiment is done, but can be based either on the knowledge of the general nature of the parameter or previously available experimental results.

In some cases, domain or expert knowledge is added to construct the prior distribution; otherwise, non-informative or vague priors can be used. When the data set is small, the latter are distributions that give little influence on the resulting posterior distribution. The posterior distribution is then summarised through summary statistics such as mean, median, mode, standard deviation. Although the philosophical interpretation of the results of the models may be different, Bayesian and frequentist results are numerically similar when the priors are vague enough, and there is a large enough quantity of data. For example, in Bayesian inference, when a non-informative prior is used, such as a flat prior, the mode of the posterior distribution is identical to the maximum likelihood estimator.

In the past, these schools of thought often conflicted, and many philosophical debates raged amongst statisticians, with many disregarding Bayesian statistics as a viable mechanism for analysis. For years Bayesian statistics failed to capture the imagination of many statisticians as the computational methods required to sample from an arbitrary probability distribution had not yet been invented. Once they were, earlier developed algorithms proved cumbersome, challenging to implement, and not widely available. However, in the last 30 years, the world has experienced the rise of the machine, and Bayesian methods have risen in popularity due to the development of fast computational algorithms. The increase in Bayesian applications has been mostly due to the adoption of Markov chain Monte Carlo (MCMC), for posterior distribution sampling [99].

Rather than subscribing exclusively to one school of thought, many statisticians now take a pragmatic approach to applied data analysis. In the cases where frequentist statistics provides more straightforward calculation, it is preferred over Bayesian models. However, in cases where frequentist statistics lacks information to arrive at a

reasonable state of inference and additional information is available, Bayesian statistics is used to address the issue. In this thesis, both frequentist and Bayesian frameworks are used for inference.

3.2 Frequentist inference

Let y be the observed data, and θ be the unknown model parameters. In a frequentist setting, inference is based on the likelihood function. The likelihood denoted by $f(y|\theta)$ acts as a function of the parameters of the model, based on the observed data. It can also be thought of as summarising the data evidence about the unknown parameter(s) θ.

The MLE is defined as:

$$\hat{\theta} \in \left\{ \underset{\theta}{\operatorname{argmax}} f(y|\theta) \right\} \tag{3.1}$$

In practice, the log-likelihood (natural logarithm) is more convenient to use, as differentiating an additive function is often easier than a multiplicative one. Furthermore, as the log-likelihood is a monotonically increasing function, it reaches the maximum value at the same point as the original function. The log-likelihood is given by:

$$\ell(\theta; y) = \ln f(y|\theta). \tag{3.2}$$

MLE is achieved by differentiating the log-likelihood with respect to θ to find it's maximum. However, one needs to check the second-order conditions as well as boundary conditions to find the absolute maximum.

In inferential statistics, the goal is to make inference or predictions about the population from which the sample is drawn. Hypothesis testing is when two mutually exclusive statements are made about a population parameter, the null and alternative hypothesis. The null hypothesis represents the status quo, or current belief about a

population parameter. Whereas, the alternative hypothesis is the claim that the population parameter has changed.

In frequentist inference, test statistics can be obtained from the data, and compared against some threshold of the distribution under the null hypothesis. The threshold is referred to as the critical value(s). It is based on a significance level α, which is the probability of a type I error, or the probability of rejecting a correct null hypothesis. If the test statistic is unlikely under the null distribution, then the null hypothesis is rejected in favour of the alternative. The data is used to construct a confidence interval, to determine possible values of the population parameter, based on a coverage probability determined by $1-\alpha$. Probability values, $p-values$, which conditions on the null hypothesis being true, is the probability of getting a result at least as extreme as the observed sample result by random chance alone. In frequentist statistics, $p-values$ are used as evidence against the null hypothesis, with low values indicating statistical significance, which suggests that the null hypothesis unlikely. However, $p-values$ should not be used as evidence alone.

If there are several frequentist candidate models to choose from, metrics such as AIC and cross-validation mean squared error can be used for model selection, which is discussed later in this chapter [81, 4].

3.3 Bayesian inference

Bayesian inference requires the likelihood function and prior distribution for the parameters within the likelihood model. A prior distribution can be formulated subjectively or objectively. A subjective approach uses historical data, past experience, or expert opinion to form the basis of the prior distribution. If a subjective approach is not possible, one can use an objective prior that is motivated by mathematical convenience, tractability, invariance properties, or to minimise the influence of the chosen prior. In simple cases, and when mathematical convenience is sought after, a conjugate prior can be used. Conjugate priors are convenient as they have the same functional form as the

likelihood, and the posterior distribution is simply an update of parameters from the prior to the posterior distribution.

In Bayesian inference, in order to make statements about the parameter of interest θ given the data y, a joint probability distribution for θ and y must be specified. The joint distribution is a product of two densities, the prior distribution $f(\theta)$ and the likelihood function $f(y|\theta)$ which yields:

$$f(\theta, y) = f(y|\theta)f(\theta). \tag{3.3}$$

The posterior density can then be derived by simply using Bayes' rule and can be calculated as such:

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{\int_\theta f(\theta)f(y|\theta)d\theta} = \frac{f(y|\theta)f(\theta)}{f(y)} \tag{3.4}$$

The denominator $f(y)$, is treated as a constant, as it is a function of the observed data y. The posterior distribution can be calculated up to proportionality, which gives the unnormalised posterior distribution:

$$f(\theta|y) \propto f(\theta)f(y|\theta). \tag{3.5}$$

However, for multi-modal likelihoods obtaining the posterior is bit more involved, but in general do not pose a problem in a Bayesian framework.

Posterior Inference

To make inference on parameters of interest, the output must be summarised to obtain features of the posterior distribution. If a point estimate is required, one can choose the mean or median of the posterior distribution, while posterior interval estimates give the range of possible values of the parameter. In this thesis, central credible intervals are used, which are intervals which contain the pre-specified proportion of the posterior

probability mass, $p\%$ and are thus bounded by the $(1-p)/2^{th}$ and the $1-(1-p)/2^{th}$ posterior distribution percentiles. Other interval estimates that are used in Bayesian inference are the highest posterior density regions (HPD), which also contains the pre-specified proportion of the posterior probability mass. Thus, any value of θ in the HPD interval has a higher posterior density than any value of θ outside of the interval. In cases where the posterior distribution is unimodal and symmetric, the central credible interval will be the same as the HPD region. The posterior probabilities of certain events of interest can also be derived from the posterior distribution and are sometimes referred to as Bayesian P-values. For example in Study II, Study III, and Study IV, the posterior probability of an area having excess risk was obtained.

After model construction, it is possible to predict an unknown but observable value $\tilde{y}$. The distribution of $\tilde{y}$ is referred to as the posterior predictive distribution. It can be used to evaluate model fit and to obtain error measurements. In Study II, and Study IV, the posterior predictive distribution was constructed to see whether the model explained the observed temporal dynamics well.

$$f(\tilde{y}|y) = \int f(\tilde{y}|\theta) f(\theta|y) d\theta. \tag{3.6}$$

When the response data is missing, it is possible to treat it as a parameter to be estimated. The joint probability model with missing data is then split into two parts, the model for the underlying complete data y which includes observed and unobserved components, and the inclusion vector I. The complete data likelihood is then given by:

$$f(y, I|\theta, \phi) = f(y|\theta) f(I|y, \phi). \tag{3.7}$$

Since y is not completely observed, the observed data likelihood is:

$$f(y_{obs}, I|\theta, \phi) = \int f(y, I|\theta, \phi) dy_{miss}, \tag{3.8}$$

If the observed data and missing data are conditionally independent given θ, with covariates x available the posterior distribution can be written as:

$$f(\theta|x, y_{obs}, I) = f(\theta|x) \int \int f(\phi|x, \theta) f(y|x, \theta) f(I|x, y, \phi) dy_{miss} d\phi. \tag{3.9}$$

However, the posterior distribution only holds when the missing data can be considered ignorable so that the missing data does not depend on y. In more complicated settings, the missing data can depend on other recorded values.

In most cases, the posterior distribution is obtained numerically; thus, the calculation of the integral is not required. In cases where the estimates of the missing data are of interest, they can be obtained via the posterior predictive distribution. The posterior predictive distribution was utilised for the missing WCR beetle counts and is discussed in *Chapter 4* and *Chapter 10*.

In practice the parameters ϕ which index the missingness are characteristics of the data collection method, but in general are not of scientific interest. However in the underreporting problem in *Chapter 9*, this parameter was of interest, and we propose a probability model to describe the underlying process of the missing observations.

Bayesian hierarchical models

Bayesian hierarchical models comprises of multiple levels that contribute to the data generating mechanism. For example, in Study II to Study IV, we make use of a hierarchical model by incorporating a spatial effect, to account for spatial autocorrelation. In this study, we model the disease risk as function of known covariates, and an additional dispersion parameter that would account for spatial autocorrelation. In a hierarchical model, the parameters of the prior distributions are referred to as hyperparameters. They can also be assigned prior distributions that are referred to as hyperprior distributions [99, 16, 65]. To obtain the posterior distribution in a Bayesian hierarchical

model, we calculate the following:

$$f(\mu|y) \propto \int f(y|\mu)f(\mu|\beta,\tau)f(\beta)f(\tau)d\beta d\tau. \tag{3.10}$$

where β and τ are parameters of μ, and $f(\beta)$ and $f(\tau)$ are the hyperprior distributions.

It is common to use a directed acyclic graph (DAG) to display hierarchical models to show the linkages between hierarchies. In a DAG, the square nodes represent observed values, whereas the circles are unknown parameters. The edges show the directional relationship between the variables and parameters. The model used in Study II is displayed in *Figure 3.1*.

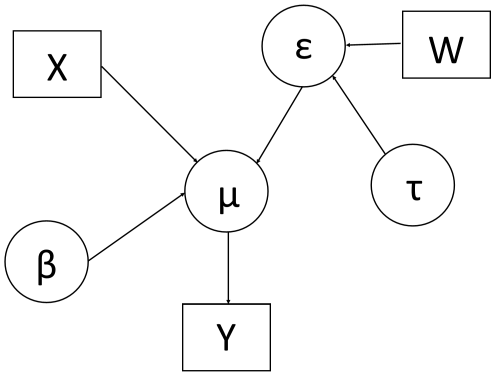


FIGURE 3.1. DAG of the fitted model in the spatiotemporal analysis of campylobacteriosis incidence (Study II).

Bayesian Computation

Bayesian inference is based on the posterior distribution, that combines the likelihood function and the prior distribution. Conjugate prior distributions are probability distributions that have the same functional form as the likelihood, so obtaining the posterior distribution is straightforward. However, for more complex multivariate or multidimensional distributions, the posterior distribution is calculated numerically, and Markov Chain Monte Carlo (MCMC) is used [99, 65].

MCMC is a class of algorithms which are used to obtain a sample from a stationary distribution, that approximates the posterior distribution. MCMC is based on the theory that samples will eventually converge to the stationary distribution, regardless of its initial state. The initial values are the starting point of the MCMC sampler, and the choice of good initial values, for example, initial values which are likely for the parameter under its posterior distribution, is known to speed up convergence. To obtain the posterior distribution or the stationary distribution $f(\theta|y)$, MCMC is run until it reaches stationarity. There are various algorithms to obtain samples from the posterior distribution, among them are the Metropolis-Hastings (MH) sampler, and the special cases of MH, the Metropolis sampler and the Gibbs sampler [65].

Let y be the observed data with likelihood $f(y|\theta)$ with the prior distribution $f(\theta)$. To obtain the posterior distribution $f(\theta|y)$ via the MH algorithm, the following steps are taken:

1. assign initial values to parameter $\theta = \theta^{(0)}$
2. propose a new parameter value $\theta^{(t)}$ from some proposal distribution $q(\cdot|\cdot)$ and calculate the acceptance ratio r:

$$r = \frac{f(\theta^{(t)}|y)q(\theta|\theta^{(t)})}{f(\theta|y)q(\theta^{(t)}|\theta)} \tag{3.11}$$

3. accept the proposed value for θ with probability $min(1, r)$
4. if the proposed value is not accepted then current state is unchanged
5. repeat steps 2-4 until convergence or set number of iterations

When the proposal distribution $q(\cdot|\cdot)$ is symmetric, then the ratio $\frac{q(\theta|\theta^{(t)})}{q(\theta^{(t)}|\theta)}$ is always equal to 1. Thus, it need not be calculated and the algorithm is referred to as the Metropolis sampler. When the conditional posterior distribution for θ is available in closed form, such as in cases where conjugate priors are used, it can also be used as the proposal distribution. The ratio r is always equal to one; and, each proposed value is accepted. This is known as the Gibbs sampler and is a special case of the MH

algorithm. The Gibbs sampler works by sampling from the conditional posterior distribution of each component of θ, while keeping the remaining variables fixed at their current state. Therefore, the key idea in Gibbs sampling is alternatively sampling from the conditional posterior distribution of each random variable.

The sample values that give rise to the posterior distribution are provided by the states of the stationary chain of the MCMC draws after a transition period referred to as burn-in.

In MCMC, thinning is the practice of discarding every k^{th} simulation draw. It can be useful when the number of samples has to be reduced for computer memory reasons, or when the posterior distribution exhibits slow or poor mixing. Issues with mixing occurs when the sample draws are autocorrelated. Poor mixing causes the sampler to explore the parameter space slowly and is also known to affect standard errors. Another way to deal with autocorrelation is the use of a non-centred reparameterisation. The aim is to separate each hierarchical layer with auxiliary variables, so each draw is independent conditional on the auxiliary parameter. An auxiliary variable is a hyperparameter that does not have direct interpretation but is introduced to improve mixing [65, 24].

In this thesis, we have followed the common practice and the convergence of the chains was visually assessed using trace plots and the marginal posterior densities of the parameters. An example of a converged chain using trace plots is given in *Figure 3.2*.

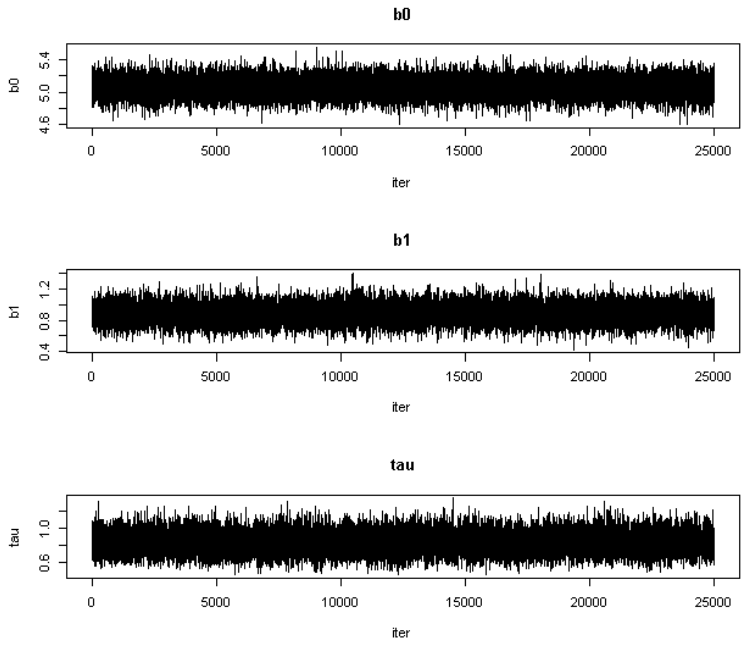


FIGURE 3.2. Examples of converged chains.

However, evaluating convergence is difficult in some situations, and visual assessment may not be appropriate. In these cases, it is worthwhile to run multiple or parallel chains with dispersed initial values. When multiple chains are run, convergence is achieved if the chains explore the same region of the parameter space. To formally test for convergence, the Gelman-Rubin (GR) convergence diagnostic can be used. The diagnostic is based on multiple chains, which evaluates MCMC convergence by estimating the difference between chain and within-chain variances for each model parameter, where large differences between them indicate non-convergence [64].

3.4 Model Comparison

To evaluate model performance and to aid in model selection, metrics such as Deviance Information Criterion (DIC) for the Bayesian models, and Akaike Information Criterion (AIC) and cross-validation error rate (CVER) for the frequentist models were used. This section will introduce them in further detail.

Deviance Information Criterion

Deviance information criterion (DIC) is used to compare the quality of fitted models against each other. It can also be as a method for model selection. In Study II and Study III, it was also used to assess evidence for spatial autocorrelation. DIC measures model performance through goodness of fit while penalising for the number of predictors used [65, 169]. DIC is given by:

$$DIC = p_D + \overline{D(\theta)} \tag{3.12}$$

where,

$$D(\theta) = -2\log\{f(y|\theta)\} + 2\log\{f(y)\}. \tag{3.13}$$

where, $D(\theta)$ is known as the Bayesian deviance, and p_D is the effective number of parameters.

$$p_D = \overline{D(\theta)} - D(\bar{\theta}) \tag{3.14}$$

In general, model fit improves as the effective number of parameters increases, so p_D acts as a penalty term. The term y are the data, θ are the model parameters, and $f(y|\theta)$ is the likelihood function. In practice, lower values of DIC indicate better model fit [65, 169].

Akaike Information Criterion

AIC, which is analogous to DIC, is often used in frequentist statistics for model comparison and selection. It is based on information theory, which aims to select the model which minimises information loss. In Study I, it was used for variable selection alongside with cross-validation, to decide which predictors should be used in the final predictive model.

AIC is defined as

$$AIC = 2k - 2\log f(y|\hat{\theta}) \tag{3.15}$$

where k is the number of parameters that require estimating in the model.

AIC takes into account goodness of fit while penalising for the number of parameters used. In the expression given above, the $2k$ acts as a penalty term to avoid overfitting.

Similar to DIC, lower values of AIC indicate a superior model. Difference of AIC values greater than or equal to three are considered to be statistically significant [81, 4]

3.5 Predictive Models

The objective of each analysis differs according to the research questions and intent of the investigator. At times, exploration of the phenomena is of importance, and quantifying how the response relates to other variables is required. In some cases, the goal of the practitioner may be purely prediction. If the objective is the latter, machine learning algorithms such as Bayesian networks can be used. If prediction and explanation of the phenomena is required, then well known statistical models such as regression can be utilised. Unlike many machine learning algorithms, the use of statistical models can quantify the association between the predictor and response, and can also provide uncertainty around estimates. Additionally if only a small number of predictors are available, it is possible that statistical models such as multinomial regression, could perform as well as the machine learning algorithms, if interaction between the variables were included.

In Study I of this thesis, the primary goal of the analysis was to find which model could best predict water quality degradation in real-time. A variety of statistical models were

employed for water quality prediction. They included a log-linear regression model, logistic regression model, discriminant analysis, regression trees, random forests and Bayesian networks. Further details of these models are given in *Chapter 7*. Model selection was based on predictive ability using cross-validation.

Cross Validation Error Rate

Cross-validation was used to evaluate model performance and for model selection in Study I. When prediction is of primary importance, the selected model must have the best out of sample performance to ensure that new or future observations are correctly predicted. Cross-validation involves splitting the data into a training set and a testing set [81]. The training data set is used to construct the model and used to obtain the prediction error for the test data set. Cross-validation can also be implemented using leave-one-out cross-validation or k-fold cross-validation. In leave-one-out cross-validation, each observation is in turn used as a test data set, and the prediction error obtained for each point. The average prediction error for a continuous response is obtained by

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^{(-i)})^2. \tag{3.16}$$

where i is the left out point, and n is the number of observations.

For classification problems, the average prediction error is calculated using:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}(y_i \neq y_i^{(-i)}). \tag{3.17}$$

When faced with a large data set, implementation of leave-one-out cross-validation can be lengthy to complete, so k-fold cross-validation can be used. In k-fold cross-validation, the data is split into k subsets, and each observation is randomly allocated to a set. Like the leave-one-out approach, each subset is used as a test data set while the remaining is used to fit the model. The model is then used to obtain the prediction

error for each subset. The cross-validation error rate for k-fold cross-validation is given by

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}(y_i - y_i^{(-i)})^2. \tag{3.18}$$

where i is the test fold and k is the number of folds.

For classification problems, the average prediction error is calculated using:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}(y_i \neq y_i^{(-i)}). \tag{3.19}$$

For classification problems, the cross-validation error rate can be displayed using a confusion matrix. Where, the diagonal entries represent the correctly classified observations, and the off-diagonals represents the misclassified points.

3.6 Spatial Models

In spatial datasets, the observations pertain to a certain point location or area. In many epidemiological applications, the data are spatially aggregated to municipal boundaries, and associations are studied via ecological regression. This technique is known as disease mapping and is a branch of spatial statistics [16]. Disease mapping techniques can highlight regional differences, prompting investigations to uncover unknown or new risk factors. They have been extensively applied to identify areas of increased disease risk [167, 168, 165, 100, 201, 191, 23].

A well-known disease mapping technique is the Besag, York and Mollie conditional autoregressive (BYM CAR) model. It was used in Study II to model the spatio-temporal distribution of campylobacteriosis incidence, and in Study III to model underreported counts in the Pennsylvania lung cancer data set. The BYM CAR model has an additional dispersion parameter that is incorporated in the ecological regression, which acts

as a smoothing parameter to handle spatial autocorrelation.

If the observations pertain to known locations and are also known to correlate with known covariates, geostatistical models such as variogram models can be utilised. The residuals of the model can be inspected to deduce whether spatial autocorrelation is present. Risk surfaces can then be interpolated using kriging. As the WCR beetle data in Study IV, were associated with geographic coordinates, a geostatistical approach was taken to account for spatial autocorrelation.

In this section, we will discuss disease mapping via the BYM CAR model, and the analysis of spatial point processes using variogram models.

Disease mapping

The Bayesian spatial conditional autoregressive (CAR) model is commonly used to study the spatial distribution of a disease and was proposed by Besag, York and Mollie in 1991 [23]. In this thesis, it was used to analyse the spatial distribution of campylobacteriosis incidence in Study II. It was also demonstrated how it could be used in underreported disease counts in Study III.

When N_i, the population at risk is known, it is natural to model the observed cases y_i, through a binomial distribution $y_i \sim Binomial(N_i, \pi_i)$. However, when the disease risk is small enough ($\pi N < 5$), or the area has a large population ($N > 50$) a Poisson approximation can be used:

$$y_i|\pi_i, N_i \sim Poisson(\pi_i N_i), \tag{3.20}$$

Where the product $\pi_i N_i$ is the expected number of cases in area i. When the population at risk N_i is small, the Poisson approximation for the binomial distribution may not

hold. The expected number of cases π_i can be modified to incorporate covariates that influence disease risk:

$$Y_i|\mu_i \sim Poisson(\mu_i), \tag{3.21}$$

where μ_i is a function of known covariates and population size N_i:

$$\log(\mu_i) = \alpha_0 + \beta^T X_i + \varepsilon_i + \log(N_i). \tag{3.22}$$

Where, α_0 represents the global risk, X_i is a vector of area-specific covariates, β is a vector of regression coefficients, N_i is the population at risk for area i, and ε_i is the area-specific random effect or spatial residual.

The CAR model incorporates a correlation structure to allow random effects to be related. Areas closer together are likely to be more similar than areas further apart, and rates are smoothed towards local or neighbourhood values. In the CAR model, a symmetric neighbourhood matrix W is defined, and the entries w_{ij} indicate whether two areas i and j are neighbours. The w_{ij} entries are usually assigned with weights equal to 1:

$$w_{ij} = \begin{cases} 1 & \text{if areas } i \text{ and } j \text{ are neighbours} \\ 0 & \text{otherwise} \end{cases} \tag{3.23}$$

Note that an area cannot be a neighbour of itself thus $w_{ii} = 0, \forall i$. Otherwise, any two areas are considered neighbours if they share a common border, or are connected by side or corner. The spatial residual ε_i is assumed to follow a conditional autoregressive

distribution defined as

$$\varepsilon_i \mid \varepsilon_{-i}, W \sim N\Big(\frac{1}{\sum_j w_{ij}} \sum_{j=1}^{N} w_{ij}\varepsilon_j, \frac{\tau_\varepsilon^2}{\sum_j w_{ij}}\Big). \tag{3.24}$$

The parameter ε_i is the spatial residual or spatial random effect for area i. The variable τ_ε represents the overall spatial precision. The local spatial precision is a proportion of the global spatial variability, weighted by the number of neighbours it has [23]. The spatial residual is the unexplained variability remaining after taking into account all the information available and relevant to the disease.

Variogram Models

Geostatistical data consists of observed values y_i, associated with a set of spatial locations z_i, within a spatial domain D. The observations are presumed to have been generated as a partial realisation of a stochastic point process. The data structure of the WCR beetle in Study IV followed this format, and so was modelled using a geostatistical model.

Let the observed counts y_i, at location z_i follow a Poisson distribution:

$$y_i|z_i \sim Poisson(\mu_i), \tag{3.25}$$

The intensity parameter μ_i is defined by

$$log(\mu_i) = \alpha_0 + \beta^T X_i + S(z_i). \tag{3.26}$$

Here, α_0 is the overall mean effect, β is a vector of regression coefficients, X_i a vector of location specific predictors, and $S(z_i)$ is the residual effect at location i. The residuals

$S(\mathbf{z})$ are assumed stationary, so that $E[S(\mathbf{z})] = 0$ and the semivariogram is defined by

$$\gamma(h) = \frac{1}{2}E[S(z) - S(z+h)]^2 \tag{3.27}$$

The variogram model assumes that the variance of $S(z)$ is constant, and spatial autocorrelation depends only on distance $h = z_i - z_j$. Therefore, we can examine the correlation from the data point pairs $\{S(z_i), S(z_j)\}$, and the sample semivariogram is calculated as

$$\gamma(h) = \frac{1}{2|N(h)|}\sum_{N(h)}(S(z_i) - S(z_j))^2 \tag{3.28}$$

Where $N(h)$ is the set of all pairwise Euclidean distance $i - j = h$, and $|N(h)|$ is the number of distinct pairs. In this case, h is a measure of distance only, but in some cases, it may be worthwhile to consider direction as well. The variogram is defined as $2\gamma(h)$ [194, 43].

The primary goal of a variogram model is to obtain the best estimates to explain the spatial autocorrelation. In simple terms, variogram models state that points become less similar with increasing distance. The majority of variogram models are described by three parameters: the nugget effect, sill and range. The nugget effect represents the variation when distance between points is zero and is supposed to reflect the measurement error. The sill is the overall variability of the spatial domain, and the range is the distance at which points are no longer autocorrelated [196, 194, 43]. *Figure 3.3* depicts a schematic diagram of a semivariogram model which shows the parameters; nugget, range and sill.

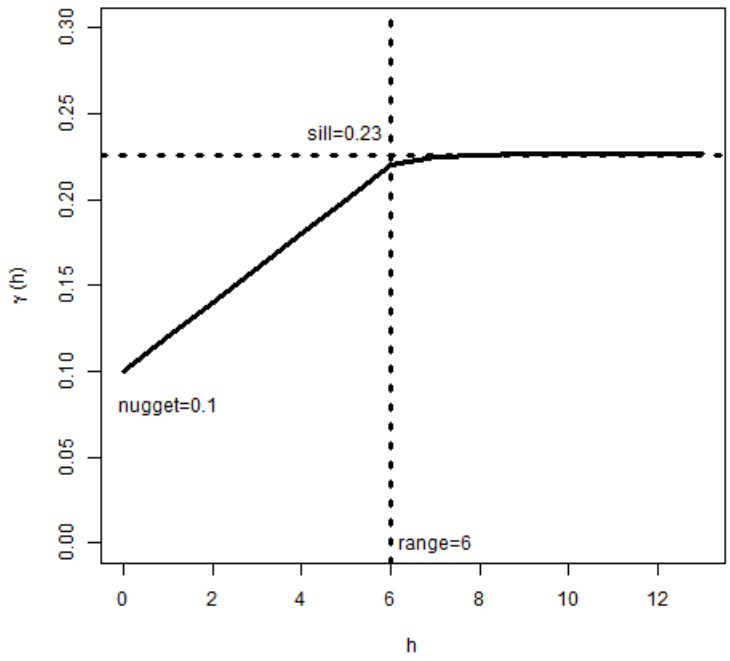


FIGURE 3.3. A schematic diagram of a semivariogram model indicating the parameters nugget, range and sill.

A suitable authorised parametric variogram model usually describes the empirical variogram. Variogram models have the following characteristics; 1) is monotonically increasing with a lag distance from the origin, 2) contains an asymptote or *sill*, 3) has a positive intercept at the origin or a *nugget*, and 4) is able to handle direction *anistropy* [194, 196]. In this thesis, we make use the powered exponential model:

$$\gamma(h) = c\left\{1 - \exp\left(\frac{h}{\theta}\right)^{\kappa}\right\}, \tag{3.29}$$

where, h is the lag distance between points, θ is a distance parameter, and κ is a scalar parameter which controls spatial smoothing. The smoothing parameter κ is constrained to lie in the interval [0,2). A schematic diagram of the powered exponential variogram model is depicted in *Figure 3.4*.

Depending on the value of κ, the powered exponential function is equivalent to other authorised variogram models [194, 43]. For example, when κ is equal to one, the powered exponential is simply the exponential variogram model, and the function approaches the sill asymptotically. In the case which $\kappa = 2$, the powered exponential is equivalent to the Gaussian variogram model, at which the semivariance near the origin slowly increases, before approaching the sill asymptotically. Variogram models can be used for interpolation (spatial prediction) via kriging or for simulation.

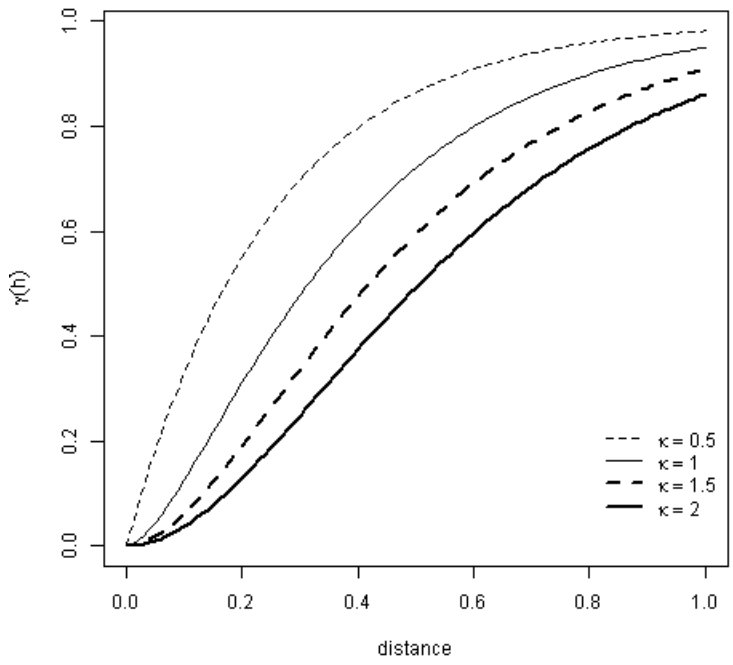


FIGURE 3.4. A schematic diagram of the powered exponential variogram model, with varying values of the smoothing parameter κ.

3.7 Temporal Models

In many data sets, observations are not only linked to location but also have a temporal aspect. The time feature marks the period at which the data were observed and collected and may play a large role in explaining the observed phenomena. For example, the average weekly temperature is affected by the season when it was collected. In this thesis, a temporal effect is evident in all data sets; for example, in Study I, we discuss how past rainfall has an impact on water quality degradation. For Study II, the reported campylobacteriosis cases corresponded to a particular year, with the

data showing abrupt changes in trend. In Study IV, we had weekly WCR beetle trap captures.

In many population biology applications, of which the WCR beetle emergence dynamics is one example, the animal emergence follow cyclical behaviour. At the beginning of the season, the number of new animals emerging are low as many have not yet matured. However, given enough time and warm days, more animals emerge, and the frequency between each new emergence is shorter. When the majority of the animals have finally matured, new emergence slows before stopping altogether.

In this section, we discuss piecewise linear regression and the Gompertz curve. Piecewise linear regression was used in Study II, the campylobacteriosis case study, to account for changes in trend and to estimate when these changes occurred. Here, the model allows for abrupt discontinuous jumps and non-differentiable changes in trend. On the other hand, the observed dynamics can often be described by some known differentiable continuous function, one of which is the Gompertz curve. In Study IV, the Gompertz function was used to model the emergence dynamics of the WCR beetle.

Piecewise Regression

In Study II, the data showed abrupt changes in incidence following regulatory changes to the poultry industry, see *Chapter 2, or Chapter 7*. Thus, the rate parameter of the Poisson distribution was modelled using a piecewise linear regression. Piecewise linear regression is known by a variety of names; such as segmented regression, broken stick model or the hockey stick model, and are characterised by a breakpoint(s) at which the trend abruptly changes. One is usually interested in the extent of the change, but the breakpoint location may also be of interest. The piecewise linear regression for the observed value Y_t at time t is defined as:

$$E[Y_t] = \beta_0 + \beta_1(t - t^*)b_1. \tag{3.30}$$

Where, β_0 is the intercept which signifies a horizontal until line until the breakpoint t^*. The term β_1, is the change in trend after the breakpoint t^*, and b_1 is a dummy variable, and $b_1 = 1$ if $t > t_1^*$; 0 otherwise. *Figure 3.5* depicts the described piecewise linear model.

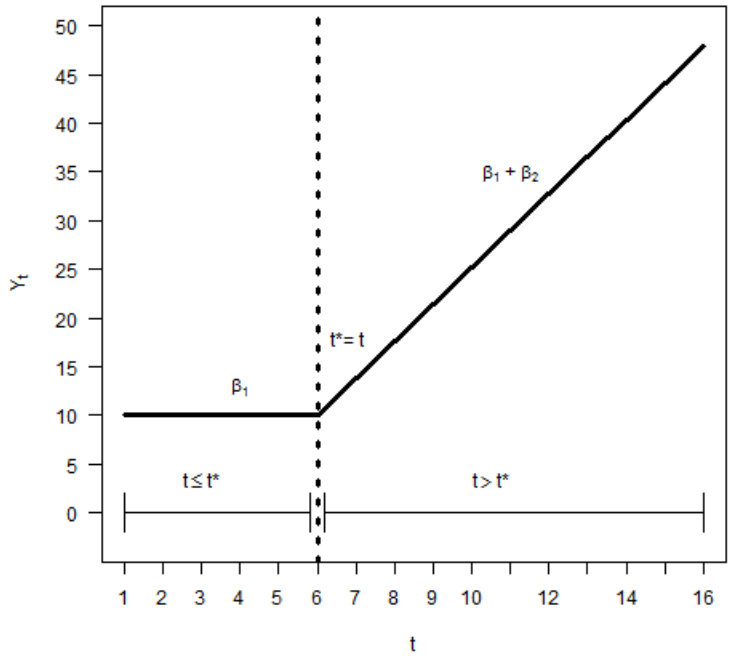


FIGURE 3.5. A schematic diagram of a piecewise linear regression model.

The breakpoints can be estimated using a grid search algorithm over the range of possible values of t. Model performance criterion such as AIC, mean squared prediction error (MSE), mean squared prediction error (MSPE), etc., can be used to choose the best breakpoint time. However, if the breakpoint is also a parameter of interest, and uncertainty measures are required for the time at which it occurred, a Bayesian approach provides an intuitive and elegant way to obtain them with an appropriate prior placed on t^*.

The Gompertz curve

Benjamin Gompertz first proposed the Gompertz curve for modelling human mortality in 1825. It is a sigmoidal function, which describes growth slowest at the beginning and the end of a given period [186, 72]. Since its inception, it has been applied to

many population biology applications [32, 3, 94, 174]. In this thesis, we make use of the well known three-parameter Gompertz function to model the observed population dynamics of the WCR beetle.

Let Y_t be the observed count for time t, which is described by the Gompertz function:

$$Y_t = \alpha \exp(-\beta \exp(-\gamma t)). \tag{3.31}$$

Where α is the upper asymptote, β is the relative starting value, and γ is the growth rate coefficient which affects the slope. To reflect the nature of population dynamics and to preserve shape, the parameters of the model are restricted to positive values so that $\alpha > 0$, $\beta > 0$, and $\gamma > 0$.

In population modelling, the time at inflection marks the peak growth and can be obtained by

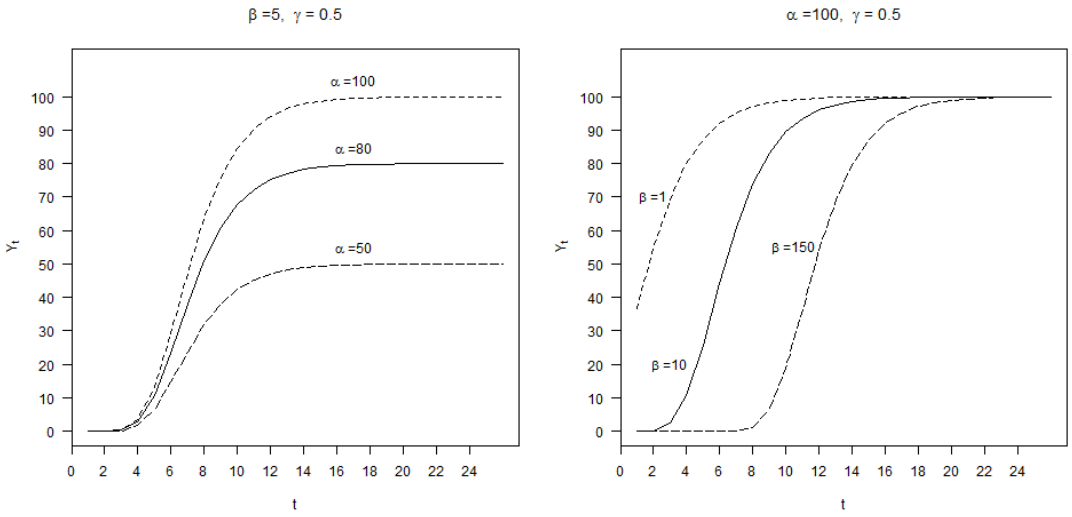
$$T^* = \frac{\log(\beta)}{\gamma}. \tag{3.32}$$

The Gompertz curve represents a cumulative function, so if we need to model incremental changes of Y_t, the derivative of the Gompertz curve can be used:

$$\frac{dY_t}{dt} = \alpha\gamma\beta \exp(-\gamma t) \exp(-\beta \exp(-\gamma t)). \tag{3.33}$$

Figure 3.6 displays how the Gompertz curve changes as one of its parameters varies. The effect of different values of the asymptote parameter α is simply a horizontal shift in the asymptote, see *Figure 3.6a*. The modification of β alters the early tail behaviour of the curve, with larger values of β having a later relative starting value, see *Figure 3.6b*. The altering of γ shows how the growth rate and slope are effected, with larger

values of γ accelerating growth which makes it reach the asymptote sooner, see *Figure 3.6c*.

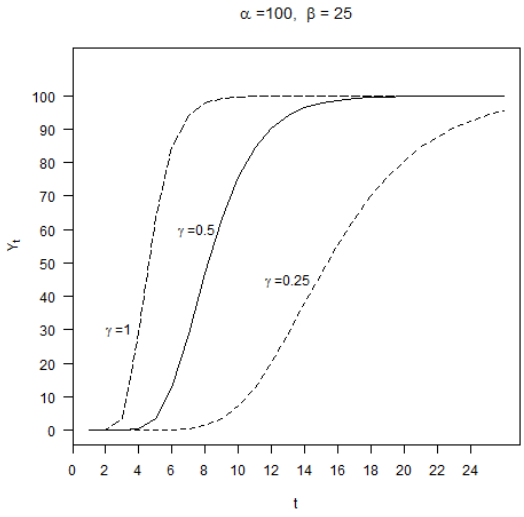


(A) varying parameter α

$\alpha = \{100, 80, 50\}$

(B) varying parameter β

$\beta = \{1, 10, 150\}$



(C) varying parameter γ

$\gamma = \{1, 0.5, 0.25\}$

FIGURE 3.6. The Gompertz curve when the parameters α, β, and γ are changed.

Data and its challenges

In a designed experiment, an investigator assigns treatments to experimental units (for example, people or geographic locations) and observes the effects of these treatments on these units. In an observational study, an investigator measures outcomes of interest for the experimental groups, but the assignment of treatment is beyond the control of an investigator.

All the studies in this thesis are observational and have come from registries. In Study I, we have information from the water quality monitoring program from a popular bathing site in New Zealand. In Study II, the records of notified campylobacteriosis cases are analysed. In study IV, analysis is based on trap counts of the western corn rootworm (WCR) beetle.

Although registries provide valuable information, they also present some problems. The purpose of the data collection is not hypothesis-driven, therefore the method of data collection may not be ideal for answering a research question. One of the problems with registry data is that unmeasured or uncontrolled variables may confound associations between exposures and outcomes [71]. The registries may also not contain all the information necessary to study the outcome of interest, so information between registries may need to be combined [137]. However, the quality of the data may vary between sources, and some key data may be missing [110]. For example, a recent audit of US Cystic fibrosis patient registries showed that therapy information and disease complications were often unrecorded [37].

In Study II, the patient information related to the campylobacteriosis cases came from health practitioners [118, 14, 125]. Over time, recording and laboratory practices can change and so affect the data collected [117]. In Study IV, the grower was responsible for counting the number of trapped beetles, and recording practices varied between locations.

All these factors need to be taken into consideration when making inference based on analysis of registry data. In this chapter, we will briefly describe the data sources used in this thesis and the challenges arising from them.

4.1 Predictions to reduce the risk of illness from freshwater swimming

The data used to build the predictive models was based on weekly water quality monitoring data of the Oreti River at Wallacetown site, for the bathing seasons of 2005/2006 to 2014/2015.

Under the Microbiological Water Quality Guidelines, a minimum of twenty water samples need to be taken from the recreational water site, and the sampling is usually conducted at weekly intervals during the bathing seasons. The water samples should be obtained from at least 30cm under the surface, in places where water depth is at least one metre, and between the hours of 8 am and 6 pm [115]. The water sample undergoes serial dilution tests, where it is used to measure *E. coli* concentration expressed as the *E. coli* most probable number (MPN) count [115]. The *E. coli* MPN count is known to be a positively-biased estimate of *E.coli* and is known to overestimate the extent of contamination [76]. The estimates for MPN are also known to be less precise than other dilution tests, such as plating and counting, and the colony-forming unit (CFU) [76]. However, compared to plating and counting, and CFU, MPN's are quicker and cheaper to obtain [195]. Therefore, it must be emphasised that the reported bacteria concentrations are only estimates of the bacterial contamination, so cannot be considered as raw

data. In study I, *E. coli* MPN was used in the analysis, as the water quality grade is defined by *E. coli* MPN count [115], and *E. coli* was the only bacterial indicator available.

In this case study, the response variable, the *E. coli* concentration and its corresponding surveillance mode were sometimes missing. In a statistical framework, there are mechanisms to deal with missing data, but the nature of the missingness must be accounted for. When there is nothing systematic that causes the data to be unobserved, it is known as missing completely at random, and ignoring the missing observations leaves inference unaffected. If data are missing systematically, then one cannot ignore the missing observations, and appropriate imputation methods have to be used. In periods of high river flow, when bathing is dangerous, sample collection may get delayed or cancelled [115]. Thus, missing data are more likely to occur during periods of rapid river flow usually proceeding heavy rainfall. To handle these instances, we proposed the use of a Bayesian network model which handles missing data by marginalising or averaging out the missing observation. This is discussed in further detail in *Chapter 7*.

As the data are based on weekly *E. coli* measurements, it is assumed that the weekly sample is representative of the entire week. However, this assumption may not hold, as water quality is known to fluctuate over space and time rapidly [21, 15]. In Study I, the response *E. coli* MPN was based on a single weekly water sample. Therefore, it does not necessarily reflect daily water quality when a person is swimming or may not represent the entire water body.

4.2 Understanding campylobacteriosis risk

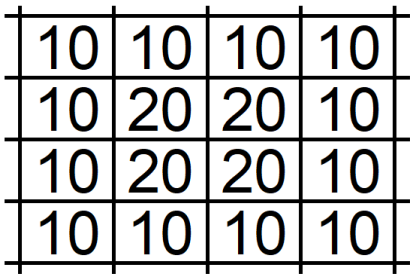
The data are based on the notified cases of campylobacteriosis from the SDHB, between the years of 2000 to 2015. The dataset was obtained from EpiSurv, which collects information on notifiable diseases from public health services [52].

Campylobacteriosis is a gastrointestinal disease that usually resolves itself within a few days. Therefore, most people rarely require a visit to the general practitioner, so as a result, many cases are unreported. Previous research has also shown that people from lower socioeconomic backgrounds may be missed in the reporting cases, as financial constraint may make them reluctant to seek medical advice [126, 167, 68]. Unreported cases are also more likely in remote rural populations, as the travel time and distance to reach health care providers may be considerable [167, 27]. The reported cases in the campylobacteriosis case study can be considered as a function of two processes: the disease risk and detection probability. The disease risk is the probability that an individual will become ill with the disease, and the detection probability describes the chance that the infected person will be included in the disease register. For the reasons described above, many cases of campylobacteriosis go systematically unrecorded.

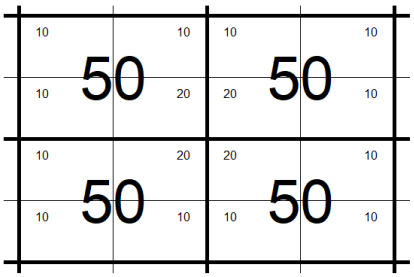
The notified campylobacteriosis cases in Study II were spatially aggregated to census area unit (CAU) level. Thus, associations between disease incidence, and the covariate socioeconomic deprivation index were investigated using ecological regression. The use of ecological regression can be controversial, as it is subject to ecological bias. Ecological bias is a fallacy in the interpretation of statistical results. Where, it is often assumed that the association at the aggregated level will transfer to the individual level [75]. The ecological bias can cause misleading results, where the association between the covariate and response can disappear or reverse at the individual level. Ecological bias is a well-known issue in disease mapping. Thus, when associations are found a cohort-based study or a case-control study often follows to investigate whether correlations exist at individual level [104, 192, 190, 150].

The modifiable area unit problem (MAUP) occurs when the same data yield different results depending on the aggregation level. MAUP was first identified by Gehlke and Biehl in 1934 and is relevant to spatially aggregated data sets [63]. In *Figure 4.1*, we depict MAUP schematically. *Figure 4.1a* shows the true number of cases, observed in each cell. If one aggregates the data, as shown in *Figure 4.1b*, then the distribution of cases is uniform over the spatial domain. In contrast, if one was to aggregate the data

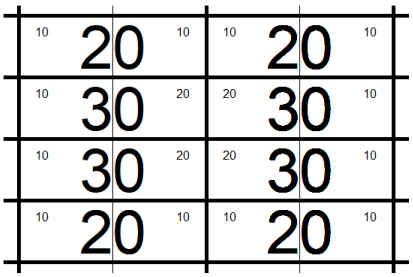
in the way shown in *Figure 4.1c*, then the central part of the spatial domain appears to have higher incidence rates than the edges.



(A) Original data.



(B) Spatial aggregation one.



(C) Spatial aggregation two.

FIGURE 4.1. The modfiable areal unit problem, displayed schematically.

The MAUP also affects the neighourhood structure that is used in models such as the BYM CAR. In panel A of *Figure 4.2*, we show how case one and case two are neighbours, and case three is neighbour of case two, but not of case one. However, if we aggregate the cells as shown in panel B of *Figure 4.2*, then case one and case three are considered neighbours. Thus, the distance between cases is probably a better measure of spatial autocorrelation. However, epidemiological data are often reported at the administrative unit level, and finer resolutions are not available.

Many municipal boundaries are arbitrarily defined by governing bodies and can be subject to change over time. For example, the analysis in Study III was completed

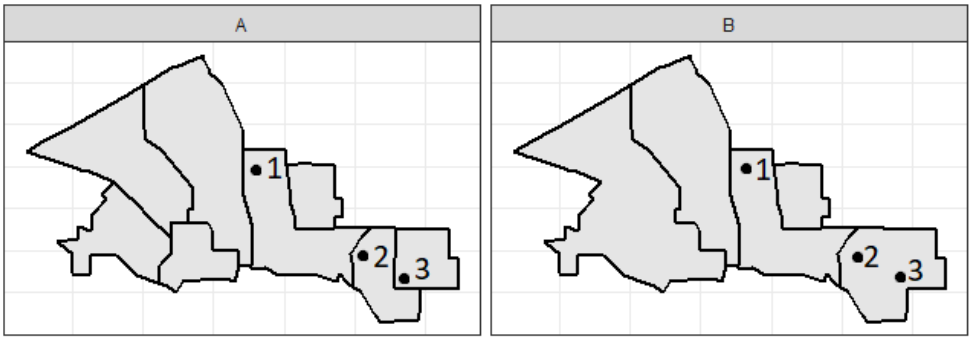


FIGURE 4.2. The effect of MAUP on neighbourhood structures.

using 2006 CAU boundaries; however, during the 2013 census, some larger CAUs in the Queenstown lakes district were split into smaller CAUs. As the SDHB is sparsely populated, the spatial distribution of population density is heterogeneous, and the CAUS vary in both areal and population size. For example, in 2006, the rural CAU of Fiordland had a population size of 18 people with the land area spanning 8287.5 km^2. In comparison, the Otago University CAU in Dunedin has a population count of 5082 but only covers $1.64km^2$. It has been suggested that analysis should be carried out on different levels of spatial aggregation, to test the robustness of the results. As this was not implemented in Study II, interpretation of the results should be made with the awareness of the above issue.

4.3 Underreporting of disease risk

In study III, a model was developed with the aim of estimating the true number of campylobacteriosis cases in New Zealand. However, as the data were only available at CAU level, suitable covariates that correlated with the disease were not available due to confidentiality reasons. Instead, we showcase the model using the Pennsylvania lung cancer data set. We have assumed a known detection probability of 0.9, and used 16 different simulated scenarios to study the performance of the model. A detection probability of 0.9 was chosen for lung cancer, as it is assumed that the majority of those afflicted by the disease are diagnosed in western countries.

4.4 Reducing the risk from western corn rootworm, (*Diabrotica virgifera virgifera*)

In Study IV, the analysis was based on traps counts of the WCR beetle during the 2014 Austrian maize growing season. The traps emitted a pheromone known to attract male WCR beetles. They were placed on maize fields or plots, where the WCR beetle had been previously sighted or expected to be seen. Thus, the traps are likely to be in areas of well-established populations, and the modelled emergence dynamics, may not represent emerging or recently established WCR beetle populations.

Most maize growers used a trap that was designed to catch 1100 to 1200 beetles. If capacity was often reached, higher capacity traps could be used, or additional traps placed. In principle, the existence of such an upper bound calls for the use of a censored distribution. However, in our data set, trap counts greater than 1200 occurred in only 0.67% of the cases. We considered this to be negligible and therefore, did not use a censored distribution.

In any data analysis, when missing data are abundant, the sample may no longer represent the population, and inference may be biased. The WCR beetle data set consisted of records in 204 trap locations, 160 of which had at least one-non zero count. The monitoring spanned for 19 weeks, resulting in 3040 observations.

The Study IV data set had many missing or blank entries. It was obvious that blanks and zeroes were used interchangeably and could represent either missing data or a genuine zero count. In most cases, such entries occurred in either the beginning or the end of the season. If missing entries happened at the beginning of the growing season, it was likely because there were no beetles to be trapped, as they were expected to be in the larvae stage. If the missing entries, occurred at the end of the growing season, then it was not known if no beetles were trapped or monitoring had ceased.

There were also 52 suspicious blanks/zeroes occurrences in the middle of the season. Because we were modelling cumulative emergence, omitting a missing data point would mean eliminating the records for the remainder of the season for the entire trap and thus losing ten traps or 6.25% of the data. Instead, after a consultation with two domain experts, we set up the following scheme:

- Any blank or missing observation until the first numeric entry was coded as zero ($n = 606$, or 19.93%)
- Any blank or zero records which occurred between two non-zero entries, at least one of which was greater than or equal to 10, were recoded as missing. ($n = 40$, or 1.31%). Otherwise, they were coded as zeroes.
- Any blank or missing records which occurred between two zero entries were coded as zeroes ($n = 2$, or 0.07%).
- Any blank or missing observations that were between at least one non zero entry were coded as missing ($n = 10$, or 0.33%).
- The trap observations were only included in the analysis until the last numeric entry, which excluded ($n = 652$, or 21.4%) observations

If an observation was deemed missing, it was estimated as a parameter in the model. After implementing the scheme above, there were 2327 non-missing observations in the data set.

In *Table 4.1*, we illustrate the scheme given above.

		Week																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Rule 1	Original Series	**blank**	**blank**	**blank**	0	2	3	10	15	36	20	28	43	29	18	blank	blank	blank	blank	blank
	Recoded Series	0	0	0	0	2	3	10	15	36	20	28	43	29	18	blank	blank	blank	blank	blank
Rule 2	Original Series	0	0	1	**NA**	2	15	36	20	28	**0**	29	18	29	18	blank	blank	blank	blank	blank
	Recoded Series	0	0	1	**0**	2	15	36	20	28	**NA**	29	18	29	18	blank	blank	blank	blank	blank
Rule 3	Original Series	0	0	0	0	2	3	0	0	**blank**	0	0	**NA**	0	0	blank	blank	blank	blank	blank
	Recoded Series	0	0	0	0	2	3	0	0	**0**	0	0	**0**	0	0	blank	blank	blank	blank	blank
Rule 4	Original Series	0	0	1	0	2	15	36	20	28	**0**	29	**blank**	29	18	blank	blank	blank	blank	blank
	Recoded Series	0	0	1	0	2	15	36	20	28	**NA**	29	**NA**	29	18	blank	blank	blank	blank	blank
Rule 5	Original Series	0	0	1	0	2	15	36	20	28	NA	29	18	29	18	**blank**	**blank**	**blank**	**blank**	**blank**
	Included in analysis	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**0**	**0**	**0**	**0**	**0**

TABLE 4.1. Demonstrating the data recoding scheme for the WCR beetle trap data

Results

5.1 Reducing the risk of illness from freshwater swimming

The objective of Study I was to determine which model could best predict water quality degradation. Here, we treated the response as both continuous and categorical by using the *E. coli* concentration (continuous), and the corresponding water quality grade (categorical). Although the methods we have proposed are well known, to our knowledge, model comparison for water quality prediction had not been thoroughly investigated in New Zealand.

The models that were explored consisted of log-linear regression, logistic regression, discriminant analysis, regression trees, random forests and Bayesian networks, for further detail, see *Chapter 7*. Model performance was assessed using leave-one-out and k-fold cross-validation.

All the models had similar cross-validation error rates. In terms of predicting acceptable (low risk) days, most models were able to estimate at least 90% of cases correctly. However, it was more important to ensure that the alert days were correctly classified as they posed the highest risk to human health. Based on predictive performance, the Bayesian network model was found to be superior. It achieved the highest classification accuracy for alert days, with 87% and 95% for leave-one-out and k-fold cross-validation respectively. From the fitted models, the next best scores for alert classifications were

from the quadratic discriminant analysis model. It achieved a classification accuracy of 75% and 86% for leave-one-out and k-fold cross-validation respectively.

5.2 Understanding campylobacteriosis disease risk

Spatial distribution of campylobacteriosis in New Zealand has been studied previously [167, 151]. However, the novelty of our research lies in using a dataset, which covered a longer and more recent period, thus allowing us to incorporate the regulatory changes in 2006.

To analyse the 2000 to 2015 notified campylobacteriosis cases in the SDHB, we used a Bayesian hierarchical model which incorporated a piecewise linear regression model and a BYM CAR normal prior to handle the spatial autocorrelation. We considered four variants of the residual normal prior, 1) no spatial autocorrelation, 2) constant spatial distribution over time, 3) spatial distribution changes over time, but the overall spatial variability does not change over time, and 4) the spatial distribution and spatial variability temporally varies. Based on DIC, the model which assumed a constant spatial effect over time was superior (Model 2). This model suggests that the spatial distribution of the disease was unchanged over the study period, and high-risk areas remained high over time.

The magnitude of the change in trend was different for urban and rural areas. *Figure 5.1* shows that the evolution of the disease differed between the two. In period 2, the decrease in notification was higher for urban areas compared to rural areas. The annual decline in notification for urban areas was 52% compared to 17% for rural areas. Therefore, risk for campylobacteriosis was higher in rural areas compared to urban ones. See *Chapter 8* for further details on the posterior estimates of the regression coefficients.

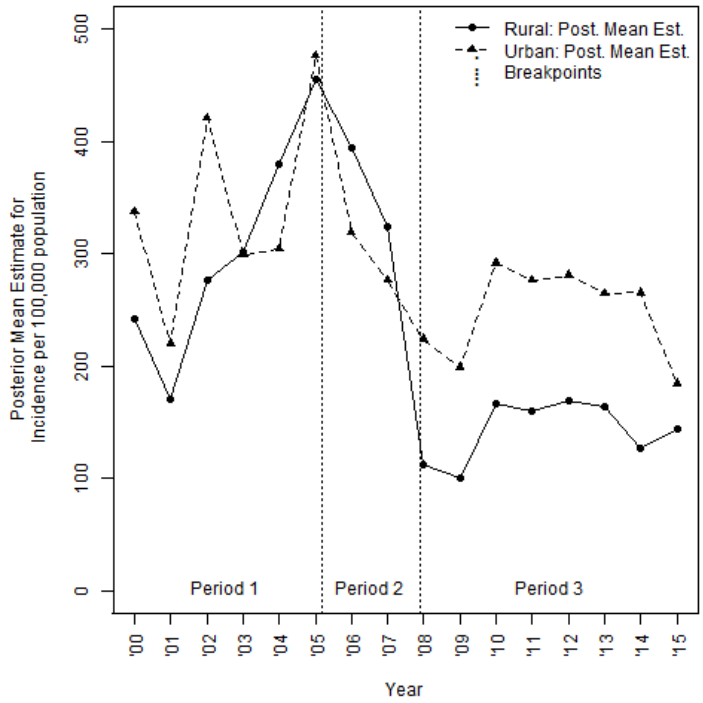


FIGURE 5.1. Posterior mean estimates for urban and rural campylobacteriosis incidence (per 100,000 populations)

5.3 Underreporting of disease risk

In Study III, we aimed to estimate the true number of cases in underreported count data and to identify which areas suffered most from underreporting.

To analyse underreported data, we developed a model in which the reparameterised Binomial distribution describes the likelihood function. This model assumes that the observed cases are a function of two processes driven by the disease risk and the detection probability, respectively. The model describes the disease risk and detection probability as functions of known covariates with a straightforward incorporation of spatial autocorrelation.

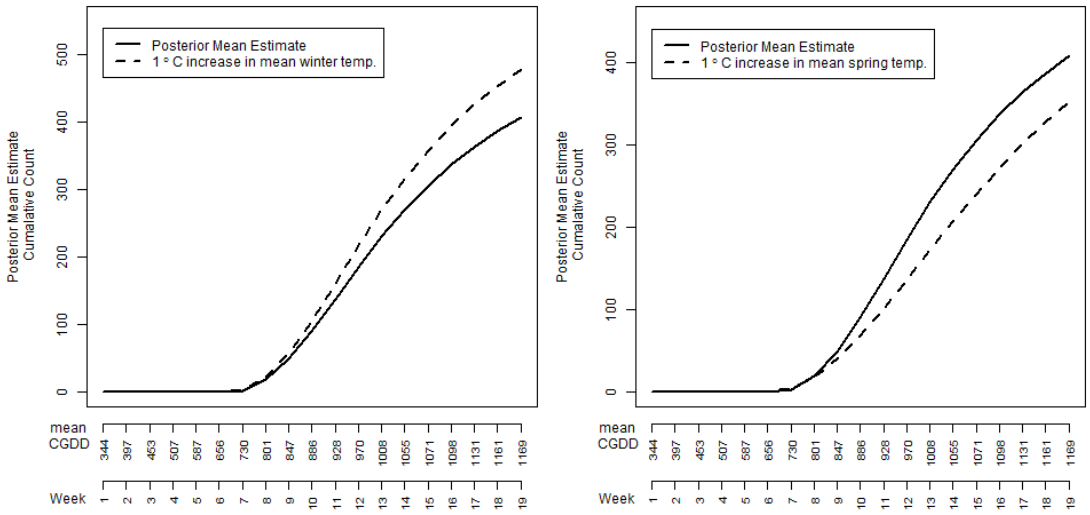
The model was applied to the Pennsylvania lung cancer data set, where we assumed that the reported lung cancer cases were underreported by 10%, and produced further 16 simulated case scenarios. Through the Pennsylvania lung cancer data set, we show how the model can be used to estimate the true number of cases, as well as uncover

areas where underreporting is severe. In the simulated scenarios, we test the model in different situations concerning the rarity of the disease and the probability of detection. The model produced posterior estimates that captured the effects of the simulated risk factors and further detail of which can be found *Chapter 9*.

5.4 Reducing the risk from western corn rootworm beetle (*Diabrotica virgifera virgifera*)

The objective of Study IV was to model the emergence dynamics of established WCR beetle populations in Austria. The three-parameter Gompertz curve was used to describe the emergence dynamics. Although the Gompertz curve is widely used in modelling the emergence dynamics of many insect populations, we treat the asymptote and growth parameters as functions of known covariates and incorporate spatially correlated errors. Through this model, we were able to quantify the effect of climatic variables on the emergence dynamics, as well as identify regions of severe WCR beetle infestation.

In Study IV, the asymptote parameter is analogous to the carrying capacity and was positively correlated with maize share and winter temperature. Therefore, higher WCR beetle populations are expected in warmer temperatures, and in areas where more maize is grown. For the growth rate coefficient, it was found that higher temperature is associated with lower growth rates, which means that the asymptote or carrying capacity is reached later. The effects of these are shown in *Figure 5.2*. From a practical perspective, this means that increased temperatures are associated with protracted WCR beetle emergence. In addition, the time of inflection is also a function of the growth rate. Therefore, warmer spring temperatures will see peak emergence occur later in the growing season.



(A) Effect of winter temperature on carrying capacity α.

(B) Effect of spring temperature on growth rate parameter γ.

FIGURE 5.2. The effect of temperature on the emergence rate

Discussion

The purpose of this thesis was to quantify the effect of environmental exposures using observational data. Unlike experimental data, collection of observational data is not hypothesis-driven, so, therefore, may not be ideal in answering research questions. However, collecting data for a specific research question can be time-consuming and costly. We thus often turn to registry data to understand what factors are associated with changes in risk.

As polluted recreational waters can lead to outbreaks of gastroenteritis or respiratory illnesses, there is a need to warn users when swimming poses a risk to human health. At the time of Study I, the local council could only advise people not to swim after a period of heavy rainfall, or when one was not able to see to their toes in knee-high water depth [51]. Although these methods provided a reasonable guideline, such practices are insufficient in understanding how different processes affected water quality degradation. Therefore, we investigated the use of various statistical models to predict water quality, to be able to use it for real-time prediction. The results of the study indicated the Bayesian network model was most suited for this purpose, as it had high predictive ability and was able to handle missing data flexibly. Bayesian network models handles missing data, by marginalising or averaging out the missing observation.

Although machine learning algorithms are adept at prediction, not all can provide estimates for the effect of a predictor, or calculate uncertainty estimates around the prediction. From a practical point of view, if the primary goal lies in prediction, this

may be unimportant. Additionally, as most Bayesian network model discretise the predictor, one can get a notion of the threshold that is associated with a change in water quality. However, as there were a small number of predictors in the study, it is possible that statistical models such as multinomial regression, could perform as well as the machine learning algorithms, if it included interaction between the variables. Since, this was not implemented in Study I, this should be further explored.

In Study II, we set out to analyse the spatio-temporal distribution of campylobacteriosis disease risk in the southern district health board (SDHB) from 2000 to 2015. Before the regulatory changes in 2006, notification rate was higher in urban centres compared to rural areas. After 2006, this reversed, and the apparent campylobacteriosis disease risk became higher for rural areas. Differences in risk are most likely due to the transmission pathways for the *Campylobacter* bacteria being different for rural populations.

A possible extension of this work, if the data was made available, is to include information on *campylobacter* contamination from different food production categories and their consumption levels. The model could then be combined with source attribution models, which would allow further insight in how campylobacteriosis risk interacts with foodborne sources and environmental exposures.

Although the risk factors for rural communities are well understood, the challenge is to minimise risk around environmental exposures. For example, if an individual lives and works on a dairy farm, one may ask what practices can be introduced so that they do not expose others to the *campylobacter* bacteria. Recommendation of such a practice is obviously beyond the scope of this thesis, but by quantifying the disparity in risk, we hope to provide useful information to policymakers. From our analysis, we found no evidence that the spatial distribution of the disease changed over time. Therefore high-risk areas remained constant over the study period. From a public health perspective, this information is useful in knowing which areas should be investigated further to understand why the changes to the poultry industry were less effective in reducing

disease risk there.

For many diseases, it is known that the reported cases underestimate the true burden of the disease. The data in Study II are a prime example of this, where it is known that the notified campylobacteriosis cases are a subset of the true number of cases. Therefore, the data are a function of two processes driven by the disease risk and detection probability, respectively. To address this, we developed a model which could handle the estimation of disease risk and detection probability. The model assumes that a covariate which correlates with the disease is available and the extent of underreporting is known. The model is non-identifiable in the intercept parameters of the disease risk and reporting probability, so elicitation of informative priors is required. However, such a strong assumption can produce biased estimates for the true number of cases. However, if a known binary variable is available, the data can be split into sub-populations, which could help address the issue of non-identifiability of some parameters [146, 78]. Another possibility is the use of capture recapture methods to estimate the detection probability, which can then be used to inform the prior distribution of detection. However, this is only possible if two disease registers are available, with some of the patients known to cross over.

The model which was developed in Study III was not implemented for estimating the true prevalence of campylobacteriosis. This was because there was not enough data on covariates which are associated with the disease.

In Europe, the focus for the invasive pest, western corn rootworm (WCR) beetle, has shifted from eradication to population control. Thus in Study IV, we modelled the emergence dynamics of established WCR beetle populations. By understanding how climatic variables interact with the emergence dynamics, we can predict timings of peak emergence. We also wished to quantify how warmer temperatures affect the carrying capacity of the population, and to provide insights on how long new beetle emergence will last.

However, the data set was rife with problems. First, there was no protocol in data recording, where it was unknown whether blank entries meant missing or zero counts. The next issue was that the sampling locations were non-random, and were chosen because the WCR beetle had been previously sighted. Such a sampling scheme was probably borne out of convenience. As the onus lay with the growers to place the traps and count the captures each week. Ideally, the traps should have been systematically laid, with the frequency of checking and emptying of traps synchronised between traps.

To our knowledge, spatially correlated residuals of the Gompertz curve parameters had not been implemented before. Therefore, to assess the feasibility of the proposed model, we used one year of WCR beetle trap captures. We now plan to expand the analysis to other years, which will allow us to assess the spatio-temporal distribution of the emergence dynamics.

Predictions to reduce risk of illness from freshwater swimming

7.1 Original publication I: Evaluating statistical model performance in water quality prediction

Avila, Rodelyn, Beverley Horn, Elaine Moriarty, Roger Hodson, and Elena Moltchanova. "Evaluating Statistical Model Performance in Water Quality Prediction." *Journal of Environmental Management* 206 (2018): 910–19. https://doi.org/10.1016/j.jenvman.2017.11.049.

Journal of Environmental Management 206 (2018) 910–919

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

Evaluating statistical model performance in water quality prediction

Rodelyn Avila [a, b, *], Beverley Horn [b], Elaine Moriarty [b], Roger Hodson [c], Elena Moltchanova [a]

[a] School of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand
[b] Institute of Environmental Science and Research, ESR, PO Box 29181, Christchurch 8540, New Zealand
[c] Environment Southland, Private Bag 90116, Invercargill 9840, New Zealand

ARTICLE INFO

Article history:
Received 29 June 2017
Received in revised form
19 October 2017
Accepted 19 November 2017
Available online 5 December 2017

Keywords:
Water quality prediction
E. coli
Statistical models
Bayesian networks

ABSTRACT

Exposure to contaminated water while swimming or boating or participating in other recreational activities can cause gastrointestinal and respiratory disease. It is not uncommon for water bodies to experience rapid fluctuations in water quality, and it is therefore vital to be able to predict them accurately and in time so as to minimise population's exposure to pathogenic organisms. *E. coli* is commonly used as an indicator to measure water quality in freshwater, and higher counts of *E. coli* are associated with increased risk to illness. In this case study, we compare the performance of a wide range of statistical models in prediction of water quality via *E. coli* levels for the weekly data collected over the summer months from 2006 to 2014 at the recreational site on the Oreti river in Wallacetown, New Zealand. The models include naive model, multiple linear regression, dynamic regression, regression tree, Markov chain, classification tree, random forests, multinomial logistic regression, discriminant analysis and Bayesian network. The results show that Bayesian network was superior to all the other models. Overall, it had a leave-one-out and *k*-fold cross validation error rate of 21%, while predicting the majority of instances of *E. coli* levels classified as unsafe by the Microbiological Water Quality Guidelines for Marine and Freshwater Recreational Areas 2003, New Zealand. Because Bayesian networks are also flexible in handling missing data and outliers and allow for continuous updating in real time, we have found them to be a promising tool, and in the future, plan to extend the analysis beyond the current case study site.

1. Introduction

Degraded water quality can be harmful to human health. Moreover, exposure to contaminated water via recreational use including swimming can result in individual illness and community outbreaks of gastrointestinal and respiratory disease (Fewtrell and Kay, 2015; Bridle, 2014; Soller et al., 2010; Yoder et al., 2008; Prüss, 1998). A consequence of these outbreaks can put unwanted pressure on health services and lead to financial losses both to the individual households, the regional and national economy (Bridle, 2014; Hunter et al., 2009; Given et al., 2006; Gleick, 2002). For these reasons, regulatory authorities manage risk by establishing guidelines for water quality to be monitored by responsible authorities.

The microbiological quality of recreational water is monitored via the presence of indicator bacteria. Annette Pruss reviewed 37 epidemiological studies on health effects from exposure to recreational water, and found that most studies reported a positive statistically significant association between the indicator-bacteria count in recreational waters and health risk in swimmers (Prüss, 1998). For freshwater, the indicator microorganisms that correlate best with health outcomes were *Escherichia coli* (*E. coli*), which is a type of fecal coliform that is used to measure the level of pollution (Odonkor and Ampofo, 2013). The presence of *E. coli* in recreational waters indicates fecal contamination which coincides with the presence of pathogenic microorganisms. Another systematic review of over 900 studies by (Wade et al., 2003) found that *E. coli* was a more consistent predictor of gastrointestinal illness than enterococci and other bacterial indicators. Although the result was not statistically significant, they found that a log (base 10) unit increase in *E. coli* count was associated with an average 2.12 (95% CI, 0.925, 4.85) increase in relative risk in fresh water. Since *E. coli* is

* Corresponding author. School of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand.
E-mail address: rodelyn.avila@pg.canterbury.ac.nz (R. Avila).

https://doi.org/10.1016/j.jenvman.2017.11.049
0301-4797/

found in all mammal and bird faeces, higher concentrations mean an increased risk of presence of other pathogens (Sampson et al., 2006; Winfield and Groisman, 2003; Edberg et al., 2000).

To ensure the risk from recreational water is minimised for the public, many governments and groups have implemented water quality standards, such as the WHO Guidelines for Safe Recreational Water Environments (World Health Organization, 2003) and the revised European Union Bathing Water Directive 2006. These regulatory tools require recreational sites to be monitored with a minimum of one monthly sample taken during the bathing season with the results of the monitoring then disclosed to the public. The responsible government must then describe their risk management measures in relation to predictable short term pollution or abnormal events (European Parliament, 2006).

Freshwater management units (FMUs) are fresh water catchments that have been set up by New Zealand regional councils in order to set freshwater objectives and limits for freshwater quality. FMUs can be grouped according to their physical characteristics as well as their social significance, i.e. who are their main users and what purpose are they used for (Ministry for the Environment, 2015). In New Zealand, the Microbiological Water Quality Guidelines for Marine and Freshwater Recreational Areas 2003 outlines the acceptable water quality for locations (FMU) designated for recreational use, where surveillance of water quality is carried out on a regular basis. These guidelines state the degree of surveillance required and if public disclosure of the water quality is required to be given based on a surveillance mode; Acceptable, Alert and Action (Green, Amber and Red). These modes are assigned to each location based on the reported *E. coli* concentration, see Table 2 (Ministry for the Environment, 2002). Acceptable/Green is defined to be generally safe for activities such as swimming and to continue routine surveillance. Alert/Amber means an increase in *E. coli* levels and sampling to be done on a daily basis and to refer to the Catchment Assessment Checklist (CAC), which is included in the aforementioned guide, to assist in identifying possible location(s) of sources of fecal contamination. Action/Red means that high levels of *E. coli* have been found and there is an increased risk to infection. The associated action plan for mode Alert/Red required to be undertaken follow the same steps as Alert/Amber with the addition of a sanitary survey with a report on sources of contamination, warning signs erected and public disclosure of a public health problem. Hence, it is especially important to distinguish Red days from the others.

Given the importance of recreational water quality, it is important not only to monitor it, but also to predict it. This is to ensure that the public can be given a timely warning of the possible contamination and the ensuing disease burden and economical loss can be avoided. This task is complicated by the fact that the water quality is influenced by a variety of factors such as seasonal changes, land-use, human activities, and extreme weather events (Kang et al., 2010; McDowell and Wilcock, 2008; Muirhead et al., 2004, 2006). It is also somewhat complicated by defining the optimal decision, and looking for a balance between false positives (warning of contamination when there is none) and false negatives (failing to spot contamination). The cost of misclassifying mode Green into Amber or Amber into Green is not as severe as these modes allow for recreational activities to be carried out. However, the misclassification of Red into Amber or Red into Green etc. should be treated seriously as it can result in severe illness.

In the past, a variety of statistical models have been used to predict water quality. Regression trees have been used to predict bathing suitability throughout Scotland (Stidson et al., 2012), and by Džeroski et al. (2000) for water quality prediction in Slovenian rivers. Discriminant analysis has been used to evaluate the spatial and temporal variations of water quality in the Gomti River, India Singh et al., 2004, and similarly in the Fuji River Basin (Shrestha and Kazama, 2007). Bayesian networks have also been used in water quality management: Ha and Stenstrom 2003 used a Bayesian network to identify the origins of storm water based on land use; and by Donald et al. (2009) to determine the risk of gastroenteritis from recycled water. The use of multiple regression models have also shown that heavy rainfall increases pollutant load (Maniquiz et al., 2010) and urban areas tend to decrease downstream water quality (Mallin et al., 2016). Moreover, Thoe et al., 2014 wanted a model to predict water quality at Santa Monica Beach that would perform better than the naive model that was used at the time. They compared model performance between five statistical models; multiple linear regression, logistic regression, partial least squares regression, artificial neural networks and classification tree and found that the all the statistical models performed better than the existing method.

The objective of this study was to find a model that could predict future *E. coli* counts or water quality modes based on preceding data in the same season or year. This prediction would be based on past values of *E. coli* counts, accumulated rainfall of a monitored upstream site in the past 48 h and river flow. The results of this study provides a basis for model suitability for real time prediction for bathing sites across Southland, New Zealand. The proposed model should be able to correctly identify mode Red days or predict higher levels of *E. coli* concentrations. An additional benefit would ideally show how the inputs and their varying levels affect water quality. This could aid in policy decisions and allow the public to better asses the level of risk in regards to recreational water use. In this case study, we apply a variety of statistical models, including log-linear regression model, logistic regression model, discriminant analysis, regression trees, random forests and Bayesian networks to predict water quality for the summers 2005–2014 for the Oreti river in Wallacetown, which is a recreational water site situated in Southland, New Zealand. The response variable, *E. coli* concentration, is treated both, as continuous counts and as categorical variable with modes Green, Amber and Red. The predictive power of each model is assessed using cross-validation and conclusions are drawn about the best practice.

2. Study site and data

The study site is situated on the Oreti River in Wallacetown, Southland New Zealand (see Fig. 1). The Oreti river in Wallacetown is a location which is identified as being of value for recreational use and is known to experience degraded water quality (Environment Southland, 2010; Environment Southland and Te Ao Marama Inc, 2010). The land use surrounding the area consists of dry stock (42%), natural state (32%), dairy farming (18%), forestry (7%) and other uses (1%). In addition, the Winton WWTP processes wastewater from the small town of Winton, the discharge is into a tributary of the Oreti River, the Winton Stream which is approximately 6 km upstream of the confluence and 23 km up stream of the Wallacetown monitoring site (Pearson and Couldrey, 2016).

These observations are for the summer months between December and April when recreational use is expected to occur see Table 1. There is variation in sample size (n) between years due to occasional missing weekly measurements. As water quality mode is derived directly from the *E. coli* counts, we can either model the reported *E. coli* concentration or the corresponding mode. These modes and their cut-off points are given in Table 2.

The data set consists of weekly measurements of *E. coli* MPN counts based on a single sample, water quality mode which is derived from *E. coli*, river flow (m^3/s) and rainfall data (mm). The *E. coli* counts were calculated using the Quantitray MPN method

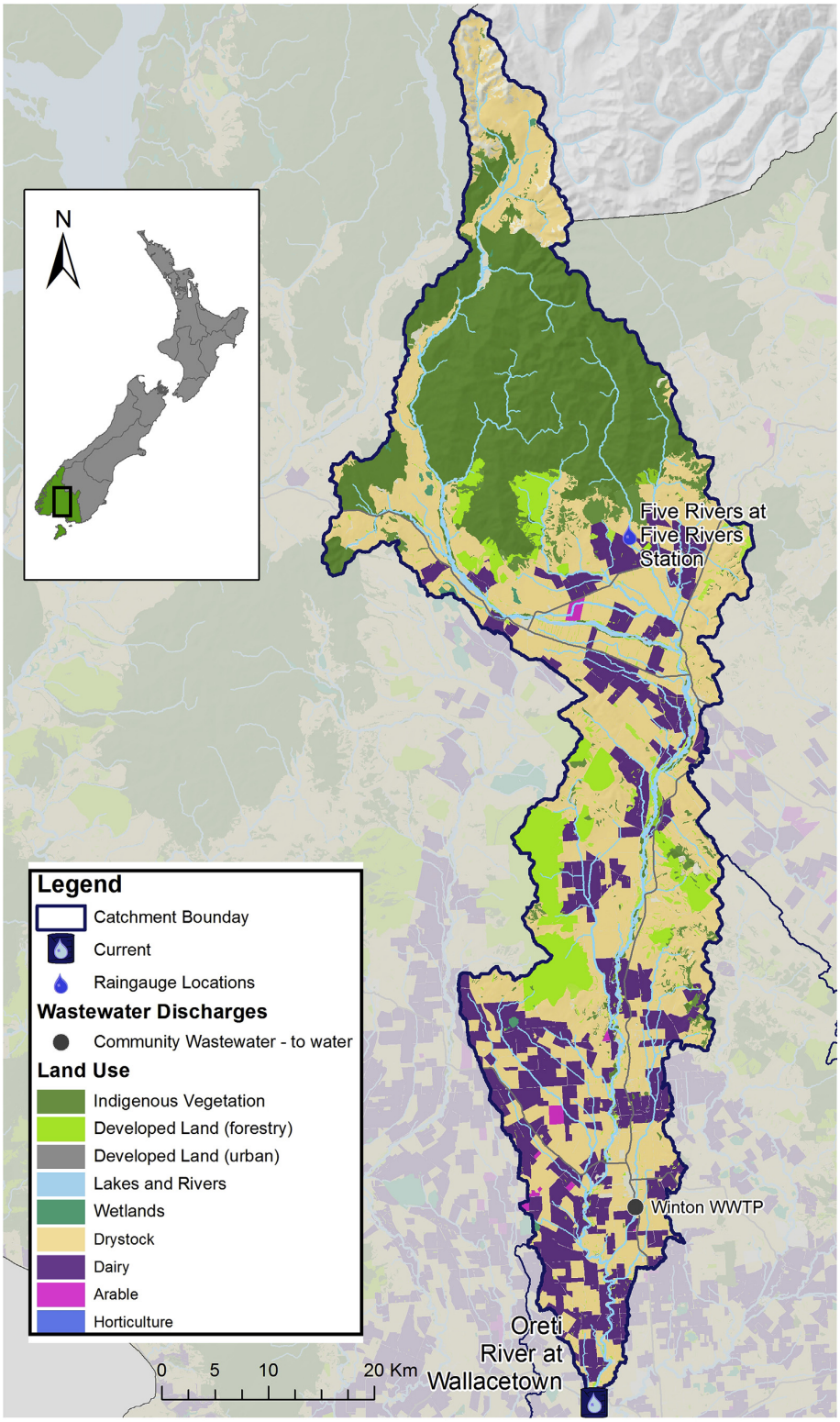


Fig. 1. A map of the Oreti River showing the study site of Wallacetown, rainfall gauge location (Five Rivers Station), waste water treatment plant (Winton) and surrounding land use.

Table 1
Frequency distribution of weekly observations of recreational water quality in the bathing seasons of 2005–2014. The modes range from Green for generally safe recreational use to Red for increased risk of infection.

Year	Mode Observed			*n*
	Green	Amber	Red	
2005–2006	10	4	4	18
2006–2007	10	3	2	15
2007–2008	13	4	1	18
2008–2009	16	2	1	19
2009–2010	14	2	1	17
2010–2011	11	0	5	16
2011–2012	13	2	1	16
2012–2013	13	2	0	15
2013–2014	14	2	1	17
Total	114	21	16	151

Table 2
Guidelines for water quality modes determined by *E. coli* concentrations as set by the Microbiological Water Quality Guidelines for Marine and Freshwater Recreational Areas 2003, Ministry of Environment, New Zealand.

mode	*E. coli* MPN/100 mL
Green	≤ 260
Amber	>260 and ≤ 550
Red	>550

and river flow was taken at the Oreti River in Wallacetown with the water level sensor at 2 mm accuracy measured every 10 s. The rainfall data which consisted of three different types of measurements was taken from a rain gauge from the 5 Rivers Station which is a connected upstream water body. It consisted of rainfall, past 24 and 48 h rainfall. The rainfall variable was measured when 0.5 mm of rain accumulated in a tipping bucket gauge. The past 24 and 48 h rainfall is the cumulative rainfall in the 24 or 48 h prior to the time of *E. coli* sample collection.

2.1. Methodology

It is known that the normal distribution can be used to approximate the Poisson distribution for large values of λ, where λ is the mean of the Poisson distribution. Therefore, if $X \sim Poisson(\lambda)$, then for large values of λ, $X \sim N(\lambda, \lambda)$ approximately (Peizer and Pratt, 1968; Cheng, 1949). Therefore, for this analysis, the reported *E. coli* MPN counts will be treated as continuous. In order to clearly distinguish between the situations when the water quality is modelled as a continuous variable from those when a categorical response is used, Y_t is used to denote continuous *E. coli* counts and Z_t to denote the corresponding modes, $Z \in \{G, A, R\}$, where G is green, A is amber and R is red. The continuous response models include naive model, multiple linear regression, dynamic regression and regression tree. Categorical responses were modelled using Markov chain, classification tree, multinomial logistic regression, discriminant analysis and Bayesian network. Although the water quality modes are ordered categories, ordinal multinomial logistic regression was not used as the effect of the predictors varied across the modes, therefore violating the proportional odds assumption (Agresti, 1996). The details of these methods are further described in this section 2.1.1.

The analysis was carried out in R, where the following packages were used to fit the appropriate model. For the regression and classification tree **rpart** was used (Therneau et al., 2015), with the random forest fitted via the **randomForest** package (Liaw and Wiener, 2002). The Markov chain was fitted using the **markovchain** package (Spedicato, 2015), and to fit the multinomial logistic regression **nnet** was used (Venables and Ripley, 2002a). Discriminant analysis was carried out using **MASS**(Venables and Ripley, 2002b) and the rest were implemented in base **R** (R Core Team, 2015).

2.1.1. Continuous response

2.1.1.1. The naive model. The naive model uses the logic that tomorrow will be the same as today. Here, the previous week's *E. coli* measurement and mode is used as the current week's prediction.

$$E(Y_t|y_{t-1}) = y_{t-1}. \tag{1}$$

This model provides a benchmark to judge the other models against.

2.1.1.2. Multiple linear regression and dynamic regression. In the multiple linear regression model the response is linearly related to a set of independent variables (Draper and Smith, 1998).

$$E(Y_t|X_t) = X_t\beta, \tag{2}$$

where X_t is a matrix of the observed variables and β is a vector of regression coefficients.

To introduce a dynamic aspect into the model, the past *E. coli* and flow levels can be included as well (Fabaozzi et al., 2006).

2.1.1.3. Regression trees. Tree-based methods partition the variable space into a set of rectangles, and then fit a model (in this case a simple linear regression) in each one Hastie et al., 2009. While there are issues with their inherent instability and lack of smoothness, tree based models often provide a simple yet powerful tool for modelling and prediction.

2.1.1.4. Random forest: regression. To address the instability of a single regression tree random forest's can be used. For regression, the same regression tree is fitted many times to bootstrap sampled versions of the training data and averages the result (Hastie et al., 2009).

2.1.2. Categorical response

2.1.2.1. Markov chain. The Markov Chain is similar to the naive model in a sense that the expected value of a stochastic process depends on the immediate past. The probability of moving from mode *i* to mode *j*, from one day to the next, is called a transitional probability, denoted p_{ij}, and is estimated from the data (Freedman, 1971). For the first order Markov Process, only the previous observation matters, and the predicted state at time t, given the observation at the previous moment $t-1$ is given by the mode of the conditional distribution $P(Z_t|Z_{t-1} = i)$, i.e., j^* such that $p_{ij^*} = \max_j p_{ij}$.

2.1.2.2. Multinomial logistic regression. The multinomial logistic regression is an extension of logistic regression when more than two outcomes are possible (Hastie et al., 2009). In order to ensure that the probabilities of all the possible outcomes add to one, the link function takes the following form:

$$P(Z_t = z|X_t) = \frac{\exp(\eta_{tz})}{1 + \sum_{z=2}^{Z}\exp(\eta_{tz})}, \tag{3}$$

where $\eta_{tz} = X_t\beta_z$ with $z = 2, \ldots, Z$ and the predicted outcome is $\max P(Z_t = z|X_t)$. The first category, $z = 1$, is called the baseline category, and $\eta_{tz} = 0$. As in the Markov model, the predicted state at

time t is given by the mode of the conditional distribution $P(Z_t = z|X_t)$, i.e. mode z^* such that $P(Z_t = z^*|X_t) = \max_z P(Z_t = z|X_t)$.

2.1.2.3. Discriminant analysis. Linear Discriminant Analysis (LDA) uses a linear combination of variables to distinguish between classes resulting in linear decision boundaries. The independent variables across the classes are assumed to be multivariate normal with a common variance-covariance. If the variance-covariance cannot be assumed equal, a modification known as quadratic discriminant analysis (QDA) is used instead (Hastie et al., 2009).

2.1.2.4. Classification trees. Classification trees are similar to regression trees: the variable space is partitioned into a set of rectangles and the most likely outcome is assigned to each. If the associated probability of an outcome assigned to a node is 1.0, the node is known as pure. Various statistics can be used to measure node purity, including misclassification error, Gini index, and cross-entropy of deviance (Hastie et al., 2009). The result can then be conveniently represented by a dendrogram.

2.1.2.5. Random forest: classification. The random forests for classification trees are similar to the regression random forest. However for classification problems, a committee of trees each cast a vote for the predicted class (Hastie et al., 2009).

2.1.2.6. Bayesian networks. A Bayesian network (BN) is a graphical model that encompasses probabilistic relationships amongst a set of variables (Fenton and Neil, 2012). A BN can be represented graphically by a directed acyclic graph (DAG), with the nodes corresponding to the variables of interest and arcs (directed edges) corresponding to the perceived relationships between them. The arcs thus represent the probabilistic relationship between the nodes and demonstrate the conditional dependence present in the network (Fenton and Neil, 2012). The Bayes' theorem is then applied to obtain probabilities of observing the class given a set of observed covariates. An example of the proposed model is given in Fig. 2 and illustrates how mode is affected by past rainfall in the last 48 h, river flow.

2.2. Model evaluation

All the fitted models were checked for compliance with the their corresponding assumptions. To assess predictive power, a leave-one-out and k-fold cross validation were used. The leave-one-out cross validation uses all but one observation in the data set to fit the model. The fitted model is then used to estimate the prediction error for the left out observation, and the step is then repeated for all observations (Hastie et al., 2009; Efron, 1983). The k-fold cross validation technique works similarly but removes k observations at a time (Hastie et al., 2009). In our case, the observations of each bathing season was removed to validate model performance with $k = 9$. The leave-one-out method gives an idea of how the model will perform in the long run while the k-fold cross validation tell us how different the bathing seasons are from one another.

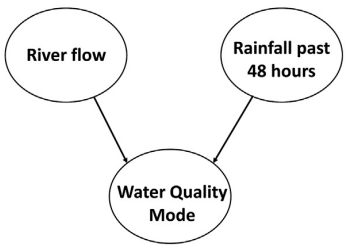


Fig. 2. Graph of the Bayesian Network used in modelling the modes observed in the Oreti River Wallacetown. The nodes are the variables and the edges show the conditional dependence between them. For example, mode is conditionally dependent on river flow.

If the continuous *E. coli* counts were being modelled, an observation was deemed predicted correctly if the estimate was within the mode boundaries of the observed mode. Model performance was evaluated via their respective cross validation error rates (CVER), i.e. the estimated proportion of misclassifications and the proportion of correct modes predicted given by the diagonal entries of the confusion matrices. The results for both leave-one-out and k-fold cross validation are reported.

Past values of weekly *E. coli* and weekly river flow were used in the dynamic regression model, with the previous two instances of each included in the model. The continuous counts of *E. coli* and river flow were log-transformed to improve compliance with various assumptions such as, for example, normality and homoscedasticity in the regression. A total of 3000 trees were constructed for the random forest (RF) for both the classification and regression. The Bayesian network model required the covariates to be discretized or split into groups. To aid in the decision of where to split the inputs, histograms of the past 48 h rainfall and river flow created to include the proportion of the observed modes at the corresponding bin. Scatter plots of river flow and past 48 h rainfall were also created with the points coloured to the corresponding mode. This visualisation allowed us to determine at which levels differentiated between modes. In addition it was found that splitting into two groups, i.e., dichotomisation, was found to be sufficient in obtaining high prediction rates. The variables were split by the following; 15.80 m^3/s and 2.00 mm for river flow and past rainfall 48 h respectively which corresponds to 60th percentile of the empirical data of these variables.

3. Results

The data contained observations of the summer months between 2005 and 2014 with an average of 17 weeks observed per year. The summary sample statistics of the variables used in the modelling process are reported in Table 3 and the predictors used for each model are given in Table 4.

The reported *E. coli* concentrations and their corresponding modes are shown in the top half of Fig. 3. The modes observed at Wallacetown were generally acceptable for recreational activities i.e. mode Amber and Green, and the poor water quality (Red) occurred only rarely. The observed modes were distributed as follows: Green 75.5%, Amber 13.9%, and Red 10.6%. Moreover, 11% of mode Green cases transitioned into Red the following week, while 80% of mode Red weeks became mode Green in the week that followed.

The 24 h accumulated rainfall is given in the bottom half of Fig. 3. It is evident that the 2005–2006 experienced more rain than other years which corresponded with a higher proportion of Amber and Red modes. The annual average rainfall for the surrounding Invercargill area is 1149 mm, with 33% falling between December to

Table 3
Table of sample statistics of variables used in modelling process, with $n = 151$.

Variable	Median	Mean	Std. dev	(2.5%, 97.5%)
E. coli MPN/100 mL	110.00	368.01	862.56	(10.00, 2832.50)
River Flow m^3/s	13.23	21.41	26.20	(4.79, 91.07)
Rainfall in mm	13.00	25.87	58.95	(0.00, 93.00)
Rainfall past 24 h in mm	0.00	3.08	6.03	(0.00, 23.12)
Rainfall past 48 h in mm	1.00	5.77	9.55	(0.00, 31.37)

Table 4
Predictors used for each model.

	Model	Predictors
Continuous Response	Dynamic Regression	River flow, rainfall past 48 Hours, *E. coli* (day previous), *E. coli* (two day previous) and river flow (day previous).
	Naive	Water quality mode (day previous).
	Regression	River flow, rainfall and rainfall past 48 h.
	Regression Tree	River flow and rainfall past 48 h.
	RF Regression	River flow, rainfall and rainfall past 48 h.
Categorical Response	Bayesian Network	River flow and rainfall past 48 h.
	Classification Tree	River flow, rainfall past 24 h and river flow (day previous).
	Linear Discriminant Analysis	River flow and rainfall past 48 h.
	Markov Chain	–
	Multinomial Logistic Regression	River flow and rainfall past 48 h.
	Quadratic Discriminant Analysis	River flow and rainfall past 48 h.
	RF Classification	River flow, rainfall and rainfall past 48 h.

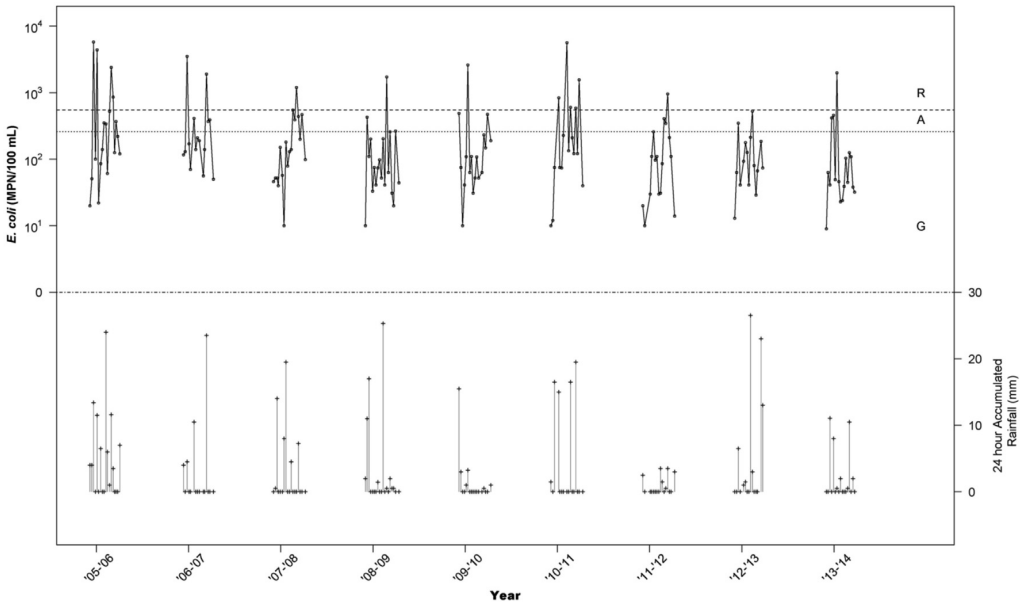


Fig. 3. Observed summer modes in the Oreti River at Wallacetown (top) and the corresponding 24 h accumulated rainfall at the Five Rivers (bottom). The boundaries of the modes are given by the horizontal lines and marked with their respective modes.

April. Rainfall is evenly distributed across the year in this area (Marcara, 2013). It can be noted that the 2005–2006 bathing season had above average rainfall was measured in the area (NIWA, 2005).

The leave-one-out cross validation results for the models are given in Table 5, with the proportion of correctly identified modes Green, Amber, Red and the cross validation error (CVER) reported. As mentioned in the methods, the naive model provides a benchmark for model performance. Overall, the naive model had the highest CVER as expected. The categorical-response models were generally better than then the continuous response models. All the models correctly predicted mode Green, with varying performance when predicting mode Red. This is with the exception of the Markov chain, as it could only correctly predict mode Green. However the prediction accuracy for the intermediate mode Amber was very poor for all models. In this study, much of the focus was to explore which model could best predict mode Red days. With this in mind the results show that the Bayesian network appeared to outperform all other models. The results for leave one out and the *k*-fold cross-validation are similar. The annual *k*-fold cross validation error rates are given in Table 6, and with the exception of the Markov chain are shown in Fig. 4. The high error rates observed for 2005–2006 summer can be explained by the fact that the proportion of Amber and Red modes were higher than other years with 44%, see Table 1. The above average rainfall that occurred at the time may account for the greater proportion of Amber and Red modes during those years (NIWA, 2005).

4. Discussion

In this study, various statistical models are used to predict water quality on a weekly basis, and their predictive accuracy is compared as well as assessing their suitability for prediction in real time. It was of particular importance to correctly predict mode Red, since it is associated with high risk of disease compared to the other modes. It was found that all models were able to accurately predict mode Green, but performed very poorly for mode Amber. For mode Red, the Bayesian network outperformed the other models, with 87%

Table 5
Model performance using leave-one-out cross validation. For proportion of correct mode predicted the closer the value is to 1 the better the performance and the value closer to zero for the cross validation error rate (CVER) indicates superior performance.

	Model	Proportion of Correct Mode Predicted for Succeeding Week			CVER
		Green	Amber	Red	
Continuous Response	Dynamic Regression	0.96	0.10	0.62	0.20
	Naive	0.77	0.24	0.06	0.38
	Regression	0.96	0.00	0.50	0.22
	Regression Tree	0.95	0.15	0.62	0.20
	RF Regression	0.95	0.14	0.62	0.20
Categorical Response	Bayesian Network	0.92	0.00	0.87	0.21
	Classification Tree	0.97	0.00	0.56	0.20
	Linear Discriminant Analysis	0.98	0.00	0.69	0.18
	Markov Chain	1.00	0.00	0.00	0.24
	Multinomial Logistic Regression	0.97	0.00	0.69	0.19
	Quadratic Discriminant Analysis	0.87	0.14	0.75	0.24
	RF Classification	0.93	0.00	0.62	0.23

Table 6
Average model performance for *k*-fold cross validation across the years. For proportion of correct mode predicted the closer the value is to 1 the better the performance and the value closer to zero for the cross validation error rate (CVER) indicates superior performance.

	Model	Proportion of Correct Mode Predicted for Succeeding Week			CVER
		Green	Amber	Red	
Continuous Response	Dynamic Regression	1.00	0.00	0.62	0.18
	Naive	0.75	0.23	0.03	0.38
	Regression	0.99	0.00	0.625	0.21
	Regression Tree	0.95	0.00	0.68	0.22
	RF Regression	1.00	0.00	0.80	0.17
Categorical Response	Bayesian Network	0.92	0.00	0.95	0.21
	Classification Tree	0.95	0.00	0.68	0.22
	Linear Discriminant Analysis	0.98	0.00	0.74	0.19
	Markov Chain	1.00	0.00	0.00	0.24
	Multinomial Logistic Regression	0.97	0.00	0.74	0.19
	Quadratic Discriminant Analysis	0.85	0.13	0.86	0.26
	RF Classification	0.94	0.00	0.71	0.22

mode Red observations correctly assigned for the leave-one-out and 95% for the *k*-fold cross validation. Therefore, we conclude that Bayesian network is the most suitable model for water quality prediction.

The water quality mode is assigned based on the reported *E. coli* MPN counts from the water body. Although other procedures to quantify *E. coli* concentration exist the Microbiological Water Quality Guidelines 2003 state that either the Membrane Filter Method, the MPN method or another accepted method must be used for *E. coli* in determining water quality (Ministry for the Environment, 2002). In our study only the MPN result was available and therefore used for analysis. However, it is important to note that the MPN is a positively-biased estimate of fecal coliform concentration (Garthright, 1997). It is also known to have wider variability in its estimates than the colony-forming-unit (CFU), another common measure of water quality, due to probabilistic basis of calculation of the MPN (Gronewold and Wolpert, 2008). Therefore, it should be noted that the phenomena modelled here may not be entirely reflective of the true and underlying process.

The ability of Bayesian networks to easily and flexibly handle missing data adds to their desirability as a modelling tool. In parametric models, missing observations are either omitted or imputed. The former may not be cost-effective: when observations are few, each one is valuable. The latter is often cumbersome, especially if the need for imputation is frequent. For example, a useful predictor for water quality is water temperature (Pratt and Chang, 2012; Carrillo et al., 1985; Faust et al., 1975). In our data set, however, it was only recorded occasionally. We have therefore excluded it from the analysis. However, in the future work, concentrating on Bayesian networks, we intend to add this variable (amongst others) to our model and investigate its effects on prediction accuracy noting its particular effect on Amber modes.

The estimation and fit of parametric models can also be affected by the presence of extreme values or outliers (Hastie et al., 2009). Bayesian networks circumvent this as the variables are discretized, thus ignoring the magnitude and influence of individual unusual observations. Furthermore Bayesian networks are also suitable for prediction in real time, as they are easily updated and there are no assumptions to check for.

However, it also known that despite its high predictive performance and aforementioned advantages, the Bayesian networks performance can be severely altered by the choice of discretization as well as the number of intervals used (Nojavan et al., 2017). In this study exploratory data analysis was used to aid where the variables should be discretized. By doing so, this allowed a better understanding in the underlying process which drives the transitional changes between modes and provides a justification of the choice of discretization. For the Bayesian network model, splitting the variables into two groups, i.e., dichotomizing, was sufficient to obtain high prediction rates. However, for other study areas, dichotomizing may not produce good enough results, and the discretization may need to be reconsidered on a site-by-site basis.

Furthermore the poor results for mode Amber demonstrate that further modelling work is required. All models performed especially poorly for the Amber mode. This may be due to the fact that it is rare (13.9% of all occurrences in our data) and transient

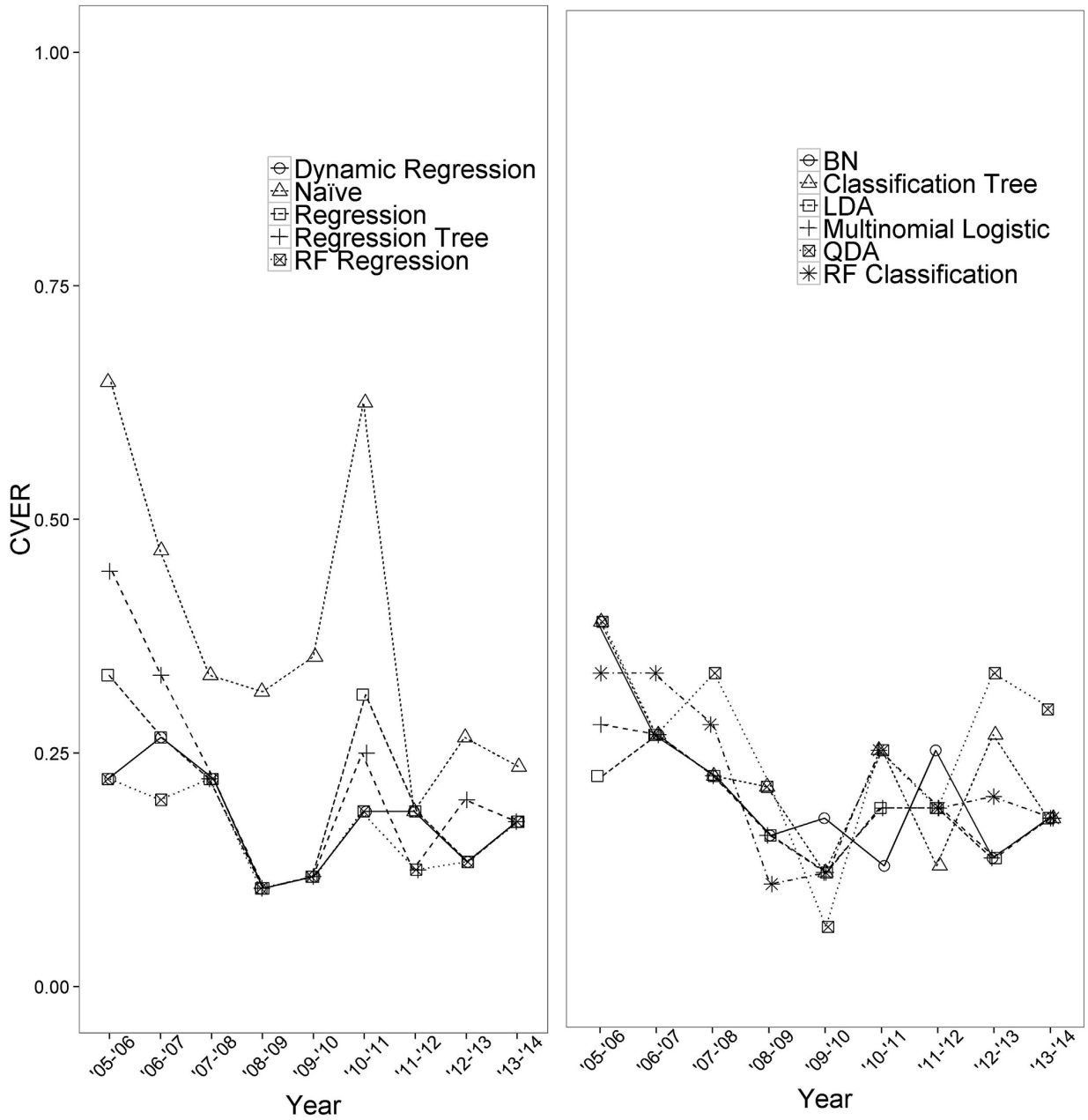


Fig. 4. Model performance by year using *k*-fold cross validation. The closer the cross validation error rate (CVER) is to zero the better the performance. Here the results for the modelling approaches are split up, the continuous response is on the left and the categorical response is on the right.

(probability of switching to another state 0.76 compared with 0.23 for Green and 0.94 for Red). Although, the transient probability is higher for Red modes, it is well established that increased rainfall corresponds to higher *E. coli* levels. However the rainfall threshold is not clearly observed or established for Amber modes. In general, the model accuracy specifically with respect to the Amber model, may be improved either by obtaining more data or by better understanding the processes behind changes in water quality and incorporation of spatial information such as land use and its proximity to the FMU as well as adding other covariates such as water temperature, electrical conductivity and other water quality indicators such as Halides.

Previous studies have explored prediction of water quality at recreational level, and evaluated changes in rivers and lakes through space and time. Deciding on an optimum model depends on the objectives of a study, with each model having its own advantages and limitations. Thoe et al., 2014 explored five statistical models; multiple linear regression, logistic regression, partial least squares regression, artificial neural networks and classification tree to predict increased levels of fecal indicator bacteria (FIB) in Santa

Monica Beach. The aim was to propose a model that would better the naive model approach that was utilised at the time. Their results showed an improvement over the naive model, with the classification tree achieving the best performance, predicting 42% of unsafe FIB levels compared to 28% by the naive. These results are consistent with our findings, as the naive and Markov chain had poor model performance.

Another example of modelling unsafe levels of fecal indicator bacteria is by Stidson et al. (2012), using regression trees they predicted 81% of unsafe levels correctly, and is the current method used for real-time water quality prediction across bathing sites in Scotland. Moreover regression trees were also used in a study conducted by De'ath et al., 2010 to evaluate the health status of the Great Barrier Reef, concluding that decreased water clarity and increased chlorophyll degrades the reef's health. In addition to high predictive power, regression trees can aid in understanding the relationship between response and predictor as the model is given by a decision tree. When modelling water quality parameters such as discharge, water temperature, dissolved oxygen etc. of Slovenian rivers, Džeroski et al., 2000 preferred the regression tree over multiple regression and the nearest neighbour method as it illustrated how the predictors affected the response. The results from our study also show that regression trees are powerful for prediction with a leave-one-out CVER of 20% and *k*-fold CVER of 22%. It also achieved one of the higher performances for mode Red, with the leave-one-out and *k*-fold cross validation yielding 62% and 68% correct respectively.

Trees are known to suffer from instability where small changes in the data can result in different partitions thus making interpretation precarious (Hastie et al., 2009). This is particularly problematic when the number of predictors are high and to address this, random forest's can be used. The number of predictors in this case study was low and therefore the results of the random forest did not differ significantly from the single classification or regression tree. For a single classification and regression tree it is easy to get an insight into decision rules if the tree is small. However for RF's this is no longer the case, as the outcome is the average result of many trees. Therefore the loss of insight of the decision rules may not be desirable when working with small data sets.

In our analysis, discriminant analysis had one of the highest performance with LDA and QDA predicting mode Red correctly with 69% and 75% respectively for leave-one-out cross validation and 74% and 86% respectively for *k*-fold cross validation. Despite the high proportion of correct mode Red predictions, for both discriminant analysis and regression trees, this level may not be high enough for policy makers and users, as the cost of false negatives can have an adverse effect on human health. Previous studies have also demonstrated Discriminant analysis' high predictive power. For instance, Shrestha and Kazama 2007 used it to model seasonal variations of water quality parameters found in surface water in the Fuji River Basin, with discriminant analysis correctly identifying 85% of the parameters variability. Moreover, Wunderlin et al., 2001 also evaluated spatio-temporal changes of water quality parameters in the Suquia River Basin, Argentina, resulting in 87% correctly predicted for temporal analysis and 75% in spatial analysis. Similarly, discriminant analysis was also fitted by Singh et al., (2004) to model spatio-temporal variations of water quality parameters in the Gomti River, resulting in 88% correctly predicted for temporal analysis and 91% in spatial analysis.

Despite their drawbacks in the choice discretization method (Gronewold and Wolpert, 2008) results of this study suggest that Bayesian networks are an ideal tool for water quality prediction as they are capable of high predictive power, see Tables 4 and 5 Like the regression tree, Bayesian networks are graphically given in a DAG, resulting in a better understanding of the relationship between the response and its predictors see Fig. 2. Other applications of Bayesian networks are by Ha and Stenstrom (2003), with their aim to differentiate between storm water origins based on land use in the Santa Monica Bay. The results of their Bayesian network correctly identified 92.3% of storm water origins. Furthermore Bayesian network's can help identify high risk groups to disease in relation to polluted water. For example Donald et al., 2009 modelled the risk of gastroenteritis associated with recycled water, with the model results indicating that the young and elderly were most susceptible to gastroenteritis. Although this is common knowledge, for other applications it shows it is capable in identifying previously unknown high risk groups. Therefore, the results from this study and those previous, suggest that a Bayesian network model should be preferred for water quality prediction as it is capable of high predictive power.

FMUs are also assigned a long term water quality grade, which is based on long term *E. coli* data trends (Ministry for the Environment, 2014). It would be of interest to investigate how BNs can be used for lakes and river grading to compare if the set limits would be similar. In addition it may be of use to evaluate how risk to GI illness differs between FMUs based on the surrounding catchment area as well as other risk factors. This can help determine if different factors unique to a site should be considered when allocating a water quality grade or if the current method is sufficient. Future modelling work will also see the Bayesian network extended to help identify possible sources of pollution and its potential in river grading.

5. Conclusion

The results from our analysis indicate that the most suitable model for real time water quality is the Bayesian network, as it could correctly predict the majority of mode red days and had a low CVER. Furthermore its ability to handle missing values, outliers and its updatability capability make it ideal for real time prediction. Future modelling work is to fit the Bayesian network model to other areas and assess its overall performance. In addition a spatial component will be included, allowing connected upstream sites and surrounding land to have an influence on the FMU, with the aim of increased accuracy of mode Amber and Red predictions. Finally, we hope that the conditional dependencies displayed in the network will aid in policy decisions regarding water quality at the recreational level.

Acknowledgements

Rodelyn Avila was jointly supported by the ESR Postgraduate scholarship and UC Connect Doctoral Scholarship (The University of Canterbury, New Zealand). Gratitude must also be extended to Environment Southland for use of their data and willingness to share their expertise and local knowledge.

References

Agresti, Alan, 1996. An Introduction to Categorical Data Analysis, vol. 312. Wiley, ISBN 0471113387.

Bridle (Heriot Watt University), Helen, 2014. Waterborne Pathogens; Detection Methods and Applications, vol. 401. Elsevier B.V, 9780444595430 (hbk.).

Carrillo, M., Estrada, E., Hazen, T.C., 1985. Survival and enumeration of the fecal indicators bifidobacterium-adolescentis and Escherichia-Coli in a tropical rain forest watershed. Appl. Environ. Microbiol. ISSN: 0099-2240 50 (2), 468–476.

Cheng, Tseng Tung, 1949. The normal approximation to the Poisson distribution and a proof of a conjecture of Ramanujan. Am. Math. Soc. 55, 396–401. https://doi.org/10.1090/S0002-9904-1949-09223-6.

De'ath, Glenn, et al., 2010. Water quality as a regional driver of coral biodiversity and macroalgae on the Great Barrier Reef Water as a driver of coral quality biodiversity regional on the Great Barrier Reef and. Ecol. Soc. Am. ISSN: 10510761 20 (3), 840–850. https://doi.org/10.1890/08-2023.1.

Donald, Margaret, Cook, Angus, Mengersen, Kerrie, 2009. Bayesian network for risk of diarrhea associated with the use of recycled water. Risk Anal. ISSN: 02724332 29 (12), 1672–1685. https://doi.org/10.1111/j.1539-6924.2009.01301.x.
Draper, N.R., Smith, H., 1998. Applied regression analysis. Technometrics. ISSN: 00359254 47 (3), 706. https://doi.org/10.1198/tech.2005.s303.
Džeroski, Sašo, Demšar, Damjan, Grbović, Jasna, 2000. Predicting chemical parameters of river water quality from bioindicator data. Appl. Intell. ISSN: 0924669X 13 (1), 7–17. https://doi.org/10.1023/A:1008323212047.
Edberg, S.C., et al., 2000. *Escherichia coli*: the best biological drinking water indicator for public health protection. In: Symposium Series (Society for Applied Microbiology). ISSN: 1467-4734, vol. 88(29), pp. 106S–116S. https://doi.org/10.1111/j.1365-2672.2000.tb05338.x.
Efron, Bradley, 1983. Estimating the error rate of a prediction rule : improvement on cross-validation. Am. Stat. Assoc. 78 (382), 316–331.
Environment Southland, 2010. Regional water plan for Southland. Visited on 09/05/2016. http://www.es.govt.nz/DocumentLibrary/Plans,policiesandstrategies/Regionalplans/RegionalWaterPlan/regional/_water/_plan.pdf.
Environment Southland, Te Ao Marama Inc, 2010. Our Health: Is Our Water Safe to Play in, Drink and Gather Kai from? Part 1 of Southland Water 2010: Report on the State of Southland's Freshwater Environment. Visited on 09/05/2016. http://www.es.govt.nz/DocumentLibrary/Researchandreports/SOEreports/water-2010-our-health.pdf.
European Parliament, 2006. Directive 2006/7/EC of the european parliament and of the council of 15 February 2006 concerning the management of bathing water quality and repealing Directive 76/160/EEC. Off. J. Eur. Commun. 64, 37–51.
Fabaozzi, Frank J., Focardi, Sergio M., Kolm, Petter N., 2006. Financial Modeling of the Equity Market from CAPM to Cointegrations, vol. 648. Wiley, New Jersey. http://vk.com/doc215711421/_317805266?hash=35172a618cc7347479/&dl=35d467177215b1d495.
Faust, Maria a., Aotaky, a. E., Hargadon, M.T., 1975. Effect of physical parameters on the in situ survival of *Escherichia coli* MC-6 in an estuarine environment. Appl. Microbiol. ISSN: 0003-6919 30 (5), 800–806. https://doi.org/10.1007/BF02090102.
Fenton, Norman E., Martin (Martin D.) Neil, 2012. Risk Assessment and Decision Analysis with Bayesian Networks, vol. 503. Taylor & Francis isbn: 9781439809105.
Fewtrell, Lorna, Kay, David, 2015. Recreational water and infection: a review of recent findings. Curr. Environ. Health Rep. ISSN: 2196-5412 2 (1), 85–94. https://doi.org/10.1007/s40572-014-0036-6.
Freedman, David, 1971. Markov Chains, vol. 382. Holden-Day, ISBN 0816230048.
Garthright, W.E., 1997. A Bayesian analysis of serial dilutions offers a worse positive bias than the MPN and proposes an inappropriate interval estimate. Food Microbiol. 14, 515–517.
Given, Suzan, Pendleton, Linwood H., Boehm, Alexandria B., 2006. Regional public health cost estimates of contaminated coastal waters: a case study of gastroenteritis at southern California beaches. Environ. Sci. Technol. ISSN: 0013936X 40 (16), 4851–4858. https://doi.org/10.1021/es060679s.
Gleick, P.H., 2002. Dirty Water: Estimated Deaths from Water-related Diseases 2000-2020. Tech. rep. Pacific Institute Research Report.
Gronewold, Andrew D., Wolpert, Robert L., 2008. Modeling the relationship between most probable number (MPN) and colony-forming unit (CFU) estimates of fecal coliform concentration. Water Res. 42, 3327–3334. https://doi.org/10.1016/j.watres.2008.04.011.
Ha, Haejin, Stenstrom, Michael K., 2003. Identification of land use with water quality data in stormwater using a neural network. Water Res. ISSN: 00431354 37 (17), 4222–4230. https://doi.org/10.1016/S0043-1354(03)00344-0.
Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome, 2009. Second. The Elements of Statistical Learning, vol. 1. Springer, ISBN 9780387848570, pp. 337–387. https://doi.org/10.1007/b94608.
Hunter, Paul R., Zmirou-Navier, Denis, Hartemann, Philippe, 2009. Estimating the impact on health of poor reliability of drinking water interventions in developing countries. Sci. Total Environ. ISSN: 00489697 407 (8), 2621–2624. https://doi.org/10.1016/j.scitotenv.2009.01.018.
Kang, Hyon, Joo, et al., 2010. Linking land-use type and stream water quality using spatial data of fecal indicator bacteria and heavy metals in the Yeongsan river basin. Water Res. ISSN: 00431354 44 (14), 4143–4157. https://doi.org/10.1016/j.watres.2010.05.009.
Liaw, Andy, Wiener, Matthew, 2002. Classification and regression by randomForest. R. News 2 (3), 18–22. http://CRAN.R-project.org/doc/Rnews/.
Mallin, Michael A., et al., 2016. Effect of human development on bacteriological water quality in coastal watersheds. Ecol. Appl. 10 (4), 1047–1056.
Maniquiz, Marla C., Lee, Soyoung, Kim, Lee-hyung, 2010. Multiple linear regression models of urban runoff pollutant load and event mean concentration considering rainfall variables. J. Environ. Sci. ISSN: 1001-0742 22 (6), 946–952. https://doi.org/10.1016/S1001-0742(09)60203-5.
Marcara, G.R., 2013. The Climate and Weather of Southland. Visited on 09/26/2017. NIWA Science and Technology Series. https://www.niwa.co.nz/sites/niwa.co.nz/files/Southland_Climate_WEB.pdf.
McDowell, Rw, Wilcock, Rj, 2008. Water quality and the effects of different pastoral animals. N. Z. Vet. J. ISSN: 0048-0169 56 (6), 289–296. https://doi.org/10.1080/00480169.2008.36849.
Ministry for the Environment, 2015. A Guide to the Ment for Freshwater Management 2014. Tech. rep. New Zealand government, Wellington.
Ministry for the Environment, 2002. Microbiological Water Quality Guidelines for Marine and Freshwater Recreational Areas. Tech. rep. New Zealand government.
Ministry for the Environment, 2014. National Policy Statement for Freshwater Management 2014. Tech. rep. New Zealand government.
Muirhead, R.W., et al., 2004. Faecal bacteria yields in artificial flood events: quantifying in-stream stores. Water Res. ISSN: 00431354 38 (5), 1215–1224. https://doi.org/10.1016/j.watres. 2003.12.010.
Muirhead, Richard William, Collins, Robert Peter, Bremer, Philip James, 2006. Interaction of *Escherichia coli* and soil particles in runoff interaction of *Escherichia coli* and soil particles in runoff. Appl. Environ. Microbiol. ISSN: 0099-2240 72 (5), 3406–3411. https://doi.org/10.1128/AEM.72.5.3406.
NIWA, 2005. National climate summary - summer 2004/05. Visited on 08/24/2016 https:/www.niwa.co.nz/sites/niwa.co.nz/files/import/attachments/sclimsum/_05/_1/_summer.pdf.
Nojavan, A., Farnaz, Qian, Song S., Stow, Craig A., 2017. Comparative analysis of discretization methods in Bayesian networks. Environ. Model. Softw. ISSN: 13648152 87, 64–71. https://doi.org/10.1016/j.envsoft.2016.10.007.
Odonkor, Stephen T., Ampofo, Joseph K., 2013. *Escherichia coli* as an indicator of bacteriological quality of water: an overview. Microbiol. Res. ISSN: 2036-7481 4 (1), 5–11. https://doi.org/10.4081/mr.2013.e2.
Pearson, L., Couldrey, M., 2016. Methodology for GIS-based land use maps for Southland. Environment Southland Publication No 2016-10. http://www.es.govt.nz/Document%20Library/Research%20and%20reports/Various%20reports/Science%20reports/Land%20use%20inputs/Report%20-%20Methodology%20for%20GIS-based%20Land%20Use%20Maps%20for%20Southland.pdf.
Peizer, David B., Pratt, John W., 1968. A normal approximation for binomial, F, beta, and other common, related tail probabilities, I. J. Am. Stat. Assoc. ISSN: 01621459 63 (324), 1416. https://doi.org/10.2307/2285895.
Pratt, Bethany, Chang, Heejun, 2012. Effects of land cover, topography, and built structure on seasonal water quality at multiple spatial scales. J. Hazard Mater. ISSN: 1873-3336 209210, 48–58. https://doi.org/10.1016/j.jhazmat.2011.12.068.
Prüss, Annette, 1998. Review of epidemiological studies on health effects from exposure to recreational water. Int. J. Epidemiol. ISSN: 0300-5771 27 (1), 1–9 doi:10. 1093/ije/27.1.1.
R Core Team, 2015. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
Sampson, Reyneé W., et al., 2006. Effects of temperature and sand on *E. coli* survival in a northern lake water microcosm. J. Water Health. ISSN: 14778920 4 (3), 389–393. https://doi.org/10.2166/wh.2006.024.
Shrestha, S., Kazama, F., 2007. Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan. Environ. Model. Softw. ISSN: 13648152 22 (4), 464–475. https://doi.org/10.1016/j.envsoft.2006.02.001.
Singh, Kunwar P., et al., 2004. Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India) - a case study. Water Res. ISSN: 00431354 38 (18), 3980–3992. https://doi.org/10.1016/j.watres.2004.06.011.
Soller, Jeffrey a., et al., 2010. Estimated human health risks from exposure to recreational waters impacted by human and non-human sources of faecal contamination. Water Res. ISSN: 00431354 44 (16), 4674–4691. https://doi.org/10.1016/j.watres.2010.06.049.
Spedicato, Giorgio Alfredo, 2015. Markovchain: Discrete Time Markov Chains Made Easy. R Package Version 0.3.1.
Stidson, R.T., Gray, C.A., McPhail, C.D., 2012. Development and use of modelling techniques for real-time bathing water quality predictions. Water Environ. J. ISSN: 17476585 26 (1), 7–18. https://doi.org/10.1111/j.1747-6593.2011.00258.x.
Therneau, Terry, Atkinson, Beth, Ripley, Brian, 2015. Rpart: Recursive Partitioning and Regression Trees. R Package Version 4.1-10. http://CRAN.R-project.org/package=rpart.
Thoe, W., et al., 2014. Predicting water quality at Santa Monica Beach: evaluation of five different models for public notification of unsafe swimming conditions. Water Res. ISSN: 1879-2448 67C, 105–117. https://doi.org/10.1016/j.watres.2014.09.001.
Venables, W.N., Ripley, B.D., 2002a. Modern Applied Statistics with S. Fourth. Springer, New York, ISBN 0-387-95457-0. http://www.stats.ox.ac.uk/pub/MASS4.
Venables, W.N., Ripley, B.D., 2002b. Modern Applied Statistics with S. Fourth. Springer, New York, ISBN 0-387-95457-0. http://www.stats.ox.ac.uk/pub/MASS4.
Wade, Timothy J., et al., 2003. Do U.S. Environmental protection agency water quality guidelines for recreational waters prevent gastrointestinal Illness? A systematic review and meta-analysis. Environ. Health Perspect. 111 (8), 1102–1109 doi:10. 1289/ehp.6241.
Winfield, M.D., Groisman, E.A., 2003. Role of nonhost Environments in the lifestyles of Salmonella and *Escherichia coli*. Appl. Environ. Microbiol. ISSN: 0099-2240 69 (7), 3687–3694. https://doi.org/10.1128/AEM.69.7.3687-3694.2003.
World Health Organization, 2003. Coastal and Fresh Waters". Geneva 1:219. Guidelines for Safe Recreational Water, vol. 1. http://www.who.int/water/_sanitation/_health/bathing/srwe2full.pdf.
Wunderlin, Daniel Alberto, et al., 2001. Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquia River basin (Cordoba-Argentina). Water Res. 35 (12), 2881–2894.
Yoder, Jonathan S., et al., 2008. Surveillance for waterborne disease and outbreaks associated with recreational water use and other aquatic facility-associated health events United States, 2005-2006. Visited on 09/06/2016. http://www.cdc.gov/mmWR/preview/mmwrhtml/ss5709a1.htm.

Understanding campylobacteriosis risk

8.1 Original publication II: Spatio-temporal analysis of differences in campylobacteriosis incidence between urban and rual areas in the Southern District Health Board, New Zealand

Jaksons, Rodelyn, Beverley Horn, Elaine Moriarty, and Elena Moltchanova. "Spatio-Temporal Analysis of Differences in Campylobacteriosis Incidence between Urban and Rural Areas in the Southern District Health Board, New Zealand." *Spatial and Spatio-Temporal Epidemiology* 31 (2019): 100304. https://doi.org/10.1016/j.sste.2019.100304.

Spatial and Spatio-temporal Epidemiology 31 (2019) 100304

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste

Spatio-temporal analysis of differences in campylobacteriosis incidence between urban and rural areas in the Southern District Health Board, New Zealand

Rodelyn Jaksons[a,b,*], Beverley Horn[b], Elaine Moriarty[b], Elena Moltchanova[a]

[a] School of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand
[b] Institute of Environmental Science and Research, ESR, 27 Creyke Road, Ilam, Christchurch 8041, PO Box 29181, Christchurch 8540, New Zealand

ARTICLE INFO

Article history:
Received 22 August 2018
Revised 7 August 2019
Accepted 8 August 2019
Available online 16 August 2019

Keywords:
Campylobcateriosis
CAR
Piece-wise regression
Breakpoint

ABSTRACT

The objective of this paper is to investigate differences in campylobacteriosis incidence between urban and rural areas in the Southern District Health Board of New Zealand between 2000 and 2015. The data were analysed using a Bayesian change-point model to evaluate how campylobacteriosis incidence changed over time and to see whether the dynamics differed between rural and urban areas. A conditional auto regressive error term was introduced to account for any spatial effects. The results of our analysis showed that campylobacteriosis incidence increased between 2000 and 2005, decreased between 2006 and 2008 then stabilised from 2009 onward. In addition we found that the changes in incidence were greater in urban areas than in rural ones.

1. Introduction

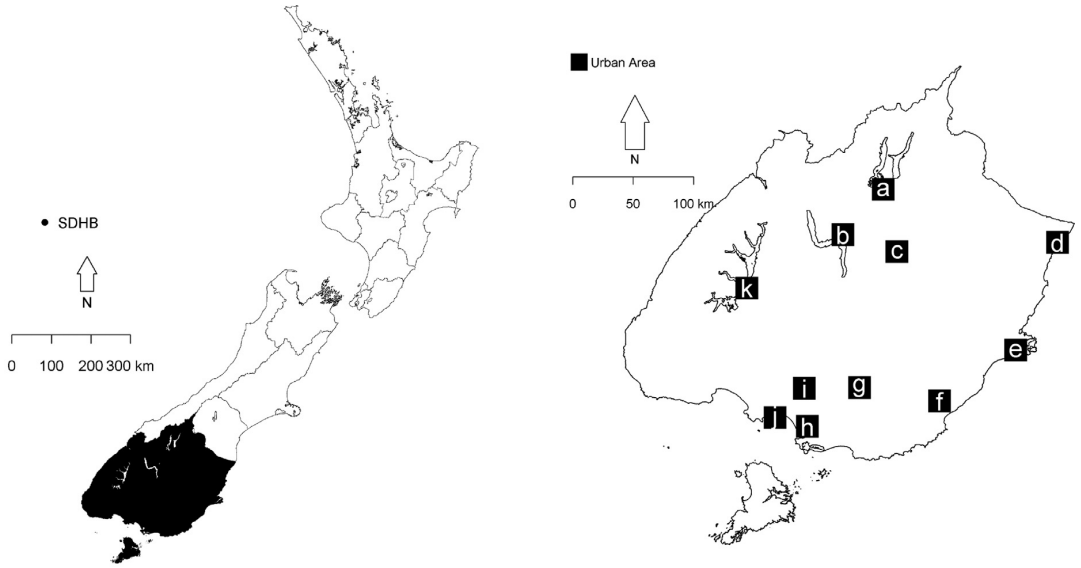
In New Zealand campylobacteriosis is a notifiable disease, where medical practitioners and laboratories must report all suspected or confirmed cases to the medical officer of health (Baker et al., 2007; Ministry of Health, 2017; New Zealand Law Resources, 1956). Starting in the 1980s until the mid 2000s, campylobacteriosis incidence steadily rose with the reported figures from New Zealand amongst the highest in the developed world. As a response, source attribution work for campylobacteriosis began in the Manawatu region in 2005 in order to discover the source of the disease source. The study indicated that more than 50% of all campylobacteriosis cases could be attributed to poultry consumption which could be linked to poultry production (Mullner et al., 2009). This resulted in the New Zealand Food Safety Authority and the Poultry Industry introducing changes in risk management strategies in 2006. This was implemented in hopes of reducing poultry associated foodborne campylobacteriosis (Sears et al., 2011). From 2007 onward campylobacteriosis notifications declined with the annual notification rate dropping from 358.8 (2002–2006) to 161.5 per 100,000 (2008). This drop in notification rates has been well documented in literature (Institute of Environmental Science and Research Limited, 2009; Pattis et al., 2017; Sears et al., 2011).

Despite the high overall notification rates, reported incidence is not uniformly spread geographically. This is especially true when examining differences in campylobacteriosis incidence between urban and rural areas. Previous research has reported that campylobacteriosis incidence is higher for rural areas compared to urban centres. One reason for this disparity could be due to a higher proportion of agricultural workers in rural areas because exposure to farm animals is believed to increase the relative risk of the disease (Gilpin et al., 2008; Lal et al., 2015; Levesque et al., 2013; Sears et al., 2011; Spencer et al., 2011b). With these factors in mind, it is reasonable to speculate that the 2006 regulatory changes to the poultry industry may have had different impacts in urban and rural areas. Although the overall changes in campylobacteriosis incidence are well documented, little is known about the spatial distribution of the disease (Institute of Environmental Science and Research Limited, 2009; Pattis et al., 2017; Sears et al., 2011).

Another attribute that must be considered is the socio-deomgraphic characteristic of an area. Previous studies have shown that notification rates were correlated with socio-demographic deprivation. As campylobacteriosis diagnosis requires a visit to the medical general practitioner, a possible explanation for differences in notification rates may be attributed to cost since people living in more deprived areas may be reluctant to seek medical advice due to financial constraints (Gillespie et al., 2008; Nichols et al., 2012; Spencer et al., 2011b).

* Corresponding author.
E-mail address: rodelyn.avila@pg.canterbury.ac.nz (R. Jaksons).

https://doi.org/10.1016/j.sste.2019.100304
1877-5845/



(a) Location of SDHB in New Zealand (b) Location of urban areas in SDHB

Fig. 1. A map of New Zealand highlighting the location of SDHB *(a)*. A map identifying urban areas *(b)*.

We employ a Bayesian spatial conditional autoregressive (CAR) model proposed by Besag, York and Mollié (Besag et al., 1991) which is a well established disease mapping tool frequently used to produce maps of underlying disease risk and to estimate the effect of possible risk factors as well as to generate new hypotheses (Besag et al., 1991; Lawson and Rotejanaprasert, 2014; Soleimani et al., 2015; Spencer et al., 2011a; 2011b; Waldhoer et al., 2008; Zhuang and Cressie, 2012). We apply it to the spatially explicit data on campylobacteriosis incidence in the Southern District Health Board (SDHB) of New Zealand in the period 2000–2015 to achieve the three following objectives; 1) to explore the differences in temporal changes in disease rates between rural and urban areas in the SDHB, 2) to investigate whether spatial dependence/spatial autocorrelation is present and to assess its variability through time, and 3) to identify disease hot spots and/or disease clusters, which can be considered areas that have high disease risk.

2. Methodology

2.1. Study area and data

This study focuses on campylobacteriosis in the SDHB of New Zealand. The SDHB provides health services to this area (see Fig. 1) (New Zealand Law Resources, 2000). The SDHB is situated in the lower South Island and covers the territorial areas of Invercargill City, Gore District, Queenstown-Lakes District, Southland District, Dunedin City, Central Otago, Clutha District and Waitaki District (Southern District Health Board). Fig. 1a shows the position of the SDHB in New Zealand. Fig. 1a and b marks the urban areas and depicts the area boundaries.

The data set consists of reported cases of campylobacteriosis from January 1, 2000 to December 31, 2015. Each reported case was either under investigation, probable, confirmed or not a case. Only confirmed or probable cases were included. A confirmed case is defined as the one for which laboratory results come back positive for campylobacteriosis.. A probable case is defined as one where the person has come in direct contact with an infected person(s) or has had contact with the same common source (Institute of Environmental Science and Research Limited, 2012; Porirua: Institute of Environmental Science Research, 2011). Cases were excluded if the person was overseas during the incubation period. For campylobacteriosis the incubation period is usually 2–5 days after source exposure (Institute of Environmental Science and Research Limited, 2012). (Fig. 2)

The reported cases were geo-coded, providing address accuracy down to census area unit (CAU) level. A CAU is a non-administrative area that is defined as an aggregation of meshblocks that is smaller in size than territorial authorities (Statistics New Zealand, b). Each CAU was assigned urban/rural profile classifications. The urban/rural classifications are used by Statistics New Zealand to distinguish between urban and rural areas at meshblock level. Urban areas can be further divided into main urban areas, satellite urban communities and independent urban communities. Rural areas can also be classified further into rural areas with high urban influence, rural areas with moderate urban influence, rural areas with low urban influence and highly rural/remote area (Statistics New Zealand, 2006a; 2006b).

For the modeling, we have considered annual CAU specific counts and the population counts were based on census data conducted in 2001, 2006 and 2013 (Statistics New Zealand, a).

The analysis was completed using 2006 boundaries so that the study area comprises of 115 urban CAUs and 94 rural CAUs, totalling to 209 CAUs. The total number of cases over the study period were 7892 and 3764 cases for urban and rural areas respectively. The average annual incidence between 2000 and 2015 for rural areas was higher compared to urban areas with 352.6 per 100,000 population and 227.7 per 100,000 population respectively. A summary of the data is given in Table 1.

The New Zealand socioeconomic deprivation index was used as a covariate in the model. It measures the area's deprivation based on census data relating to income, home ownership, employment, qualifications, family structure, housing, access to transport, and communications. The deprivation index scores range from 1 for the least deprived to 10 for the most deprived (Atkinson et al., 2014; Salmond et al., 2002; 2006).

Fig. 2. A map of the urban area boundaries in Fig. 1b.

Table 1
Summary of the data used in analysis. It summarises the number of CAUs their corresponding population (New Zealand), total number of reported cases and annual average incidence.

	No. CAU's	Population-time at risk	No. cases	Incidence per. 100,000/year
Urban	115	3,465,237	7892	227.75
Rural	94	1,067,397	3764	352.63
Total	209	4,532,634	11,656	257.16

2.2. Methodology

To investigate whether spatial autocorrelation was present, a Bayesian spatial conditional autoregressive (CAR) model was fitted (Besag et al., 1991). This model allows areas closer together to be more similar in incidence than areas further apart. By doing so the model smooths disease incidence and allows spatial effects to highlight regional differences.

Let Y_{it} denote the reported number of cases in area i at time $t = 2000, 2001, 2002..., 2015$. Y_{it} is assumed to follow a Poisson distribution with intensity μ_{it} with N_{it} denoting the respective population at risk:

$$Y_{it} \sim Poisson(\mu_{it} N_{it})$$

The Poisson intensity parameter μ_{it} is the rate of cases per person for time t in area i, and is modelled via a log-linear piecewise regression Eq. (1). Piecewise regression models are utilised when the data exhibits abrupt changes in trend (Martinez-Beneito et al., 2011; Muggeo, 2003; Tiwari et al., 2005).

$$\begin{aligned}\log(\mu_{it}) = {} & \alpha_p + \beta_{0p} t + \beta_{1p}(t - t^*_{1p}) b_{1p} \\ & + \beta_{2p}(t - t^*_{2p}) b_{2p} + \beta_{3p} x_{it} + \epsilon^*_i \end{aligned} \quad (1)$$

where t^*_{1p} and t^*_{2p} represent the unknown breakpoints thus splitting the study period into three time intervals. The rural/urban area classification is given by $p = \{rural, urban\}$. The variables t_{1p} and t_{2p} are dummy variables where $b_{1p} = 1$ if $t > t^*_{1p}$; 0 otherwise and $b_{2p} = 1$ if $t > t^*_{2p}$ and 0 otherwise. β_{0p} is the trend coefficient

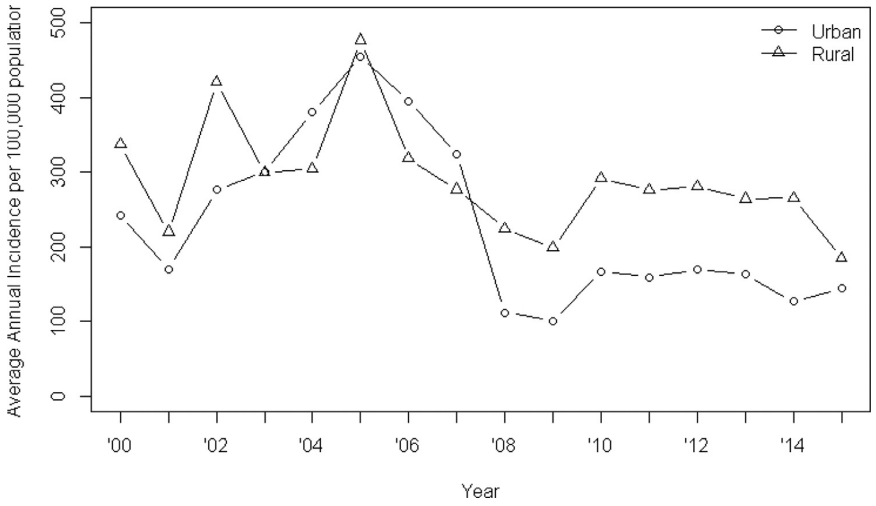


Fig. 3. Annual average incidence per 100,000 population for rural and urban areas. For both urban and rural areas incidence rose between 2000 and 2005, then fell until 2008 and is stable thereafter. The changes in urban areas appear to be greater than in rural areas.

in the first time interval, β_{1p} is the change to the trend coefficient in the second time interval and β_{2p} is the change in the trend coefficient in the third time interval. Note that these changes are additive. β_{3p} is the effect of socioeconomic deprivation and x_{it} is the deprivation score in the area i at time t.

The regression parameters α_p, β_{0p}, β_{1p}, β_{2p} and β_{3p} were assigned standard non-informative normal priors with mean 0 and precision (inverse variance) 0.001. The breakpoints were given a uniform prior where $t^*_{1p} \sim U(2002, 2013)$ and $t^*_{2p}|t^*_{1p} \sim U(t^*_{1p} + 2, 2013)$. To ensure indentifiability and model convergence the latter was assigned a lower limit to be two years after the first breakpoint.

The spatial residual ϵ^*_i is the measure of local spatial variability. We have considered four possible ways to model it. In the first case, we assume no spatial variability at all, and thus $\epsilon^*_i \equiv 0 \forall i$ Eq. (2). For the other three, we assume conditional autoregressive (CAR) prior where (1) the spatial residual does not depend on time (i.e., the way in which incidence varies geographically does not change with time, high risk locations remain high risk locations, and low risk locations remain low risk locations respectively.), (2) the spatial residual depends on time, but the overall spatial variability, as expressed via precision parameter τ remains constant Eq. (3), and (3) the spatial residual depends on time and the overall spatial variability is allowed to change as well Eq. (4). We refer to these four model as the temporal model, the common CAR model, stable CAR model and temporal CAR model respectively.

Note that because we were not able to obtain age-specific population counts for smaller CAU's due to confidentiality concerns, we were not able add age-standardization to the model.

$$\epsilon^*_i \equiv 0 \tag{2}$$

$$\epsilon^*_i \sim N(\overline{\epsilon}_{-i}, \tau m_i) \tag{3}$$

$$\epsilon^*_{it} \sim N(\overline{\epsilon}_{-it}, \tau m_i) \tag{4}$$

$$\epsilon^*_{it} \sim N(\overline{\epsilon}_{-it}, \tau_t m_i) \tag{5}$$

Note that the $\overline{\epsilon}^*_{-i}$ refers to the mean of ϵ_i in the neighbourhood of i with the area i itself always excluded. Two areas are considered neighbours if they share a common border or are connected by side or corner. The parameter τ is the inverse spatial variance or spatial precision, which is given the prior $\tau \sim Gamma(0.1, 0.1)$, and m_i is the number of neighbours for area i (Besag et al., 1991).

The posterior distributions of α_p, β_{0p}, β_{1p}, β_{2p}, γ_p, t^*_{1p} and t^*_{2p} were summarised with their corresponding 95% credible interval reported. Model selection was conducted using deviance information criterion (DIC). Models with smaller DIC values are preferred with differences greater than three indicating substantial difference (Spiegelhalter et al., 2002).

The analysis was run using WinBUGS (Lunn et al., 2000) via the **R2WinBUGS** (Sturtz et al., 2005) package in **R** (R Core Team, 2015). A total of 10,000 iterations were run with a burn-in of 5000 iterations. Burn-ins are iterations which are discarded to ensure model convergence. Model convergence was visually assessed from trace plots and marginal posterior distributions.

3. Results

3.1. Exploratory data analysis

The annual average incidence of campylobacteriosis for urban and rural is displayed in Fig. 3. Both areas experienced increased incidence between 2000 and 2005. A decrease is then observed until 2008 where reported incidence levels out. The plot also shows that the incidence decrease was larger for urban areas than rural areas.

The age-specific incidences for rural areas and urban areas (where age distribution is available) is given in Fig. 4. For rural areas, the highest reported incidence came from the 0–4 year age group followed by the 15–24 years (Fig. 4a). In urban areas, the 15–24 year age group had the highest incidence between 2000 and 2007 (Fig. 4b). This changed to the 0–4 years after 2008. The temporal pattern for both areas is similar to the overall pattern seen in Fig. 3.

The individual CAUs annual incidence over the study period is displayed in Fig. 5. The left panel shows the incidence of campylobacteriosis in individual CAUs in rural areas, whereas the right panel demonstrates the incidence of campylobcateriosis in CAUs in urban areas. The saturation of each line has been altered to reflect the respective population size. If areas follow the same temporal pattern, the lines should layer on top of each other making it darker, thus emphasising a common trend. For urban areas, a drop

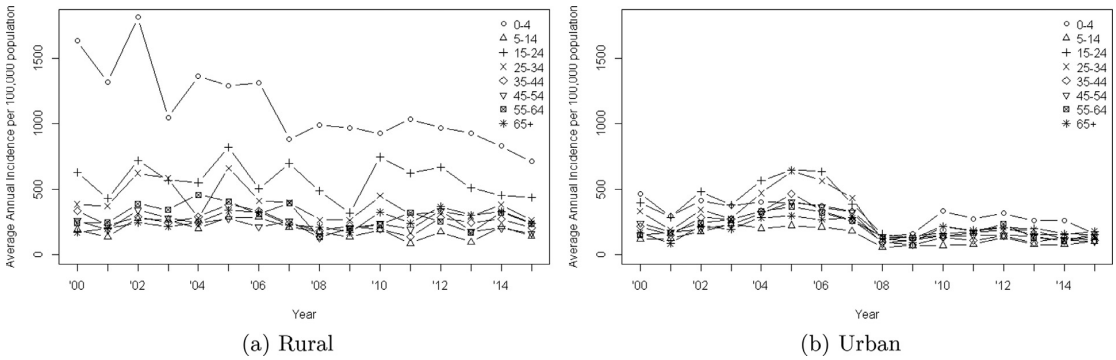


Fig. 4. Age-specific incidence for rural areas *(a)* and urban areas *(b)*. For rural areas, the 0–4 year-olds had the highest incidence. In urban areas the 15–24 year-olds had the highest incidence of campylobacteriosis between 2000 and 2007. After 2008 the 0–4 year-olds became the highest incidence group.

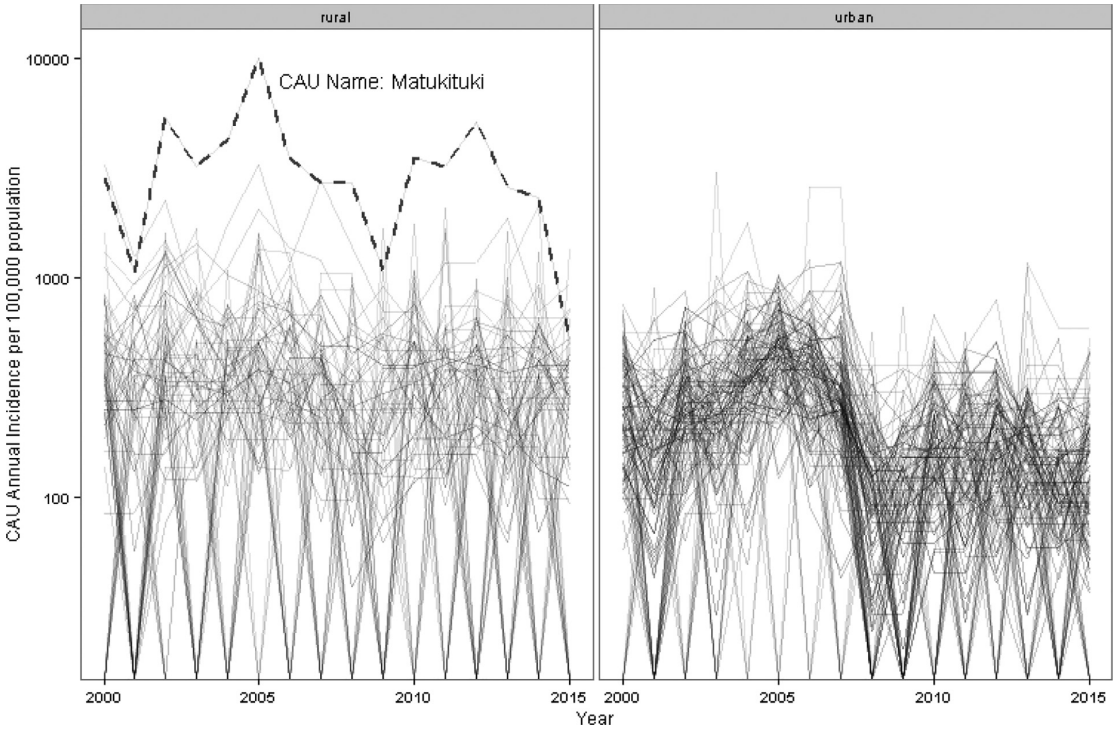


Fig. 5. Annual incidence of campylobacteriosis by CAU. The left panel shows the incidence of individual CAUs in rural areas, whereas the right panel demonstrates the CAUs in urban areas. The Matukituki CAU has been highlighted to show that it is above average for the study period.

in incidence is observed between 2006 and 2009 (Fig. 5). The time series for one rural area, Matukituki in the Queenstown Lakes District, has been made darker and given a dashed line to emphasise that it was consistently above average for the duration of the study period.

3.2. Model results

To aid in model selection the deviation information criterion (DIC) was used to assess goodness of fit, with smaller DIC values preferred (Spiegelhalter et al., 2002). The relative DICs are reported in Table 2. The best model has a DIC of zero and is used as a baseline for comparison. Table 2 shows that the spatial models perform better than the temporal model, providing evidence for the presence of spatial autocorrelation.

Table 2
The relative DIC used for model selection. The best model is the common CAR model and is given in bold. The DIC results show that the spatial models perform better than the temporal model, indicating the presence of spatial autocorrelation which has remained stable over time.

Model	Spatial residual ϵ_i^*	DIC Difference
Temporal	$\epsilon_i^* \equiv 0$	1191.4
Common CAR	$\epsilon_i^* \sim N(\overline{\epsilon}_{-i}, \tau m_i)$	**0**
Stable CAR	$\epsilon_i^* \sim N(\overline{\epsilon}_{-it}, \tau m_i)$	341.8
Temporal CAR	$\epsilon_i^* \sim N(\overline{\epsilon}_{-it}, \tau_t m_i)$	287.3

The common CAR model, which assumes that the spatial variability is present and its pattern stable over time, is the superior model with a minimum DIC difference of 287.3. The next best alternative is the temporal CAR, followed by the stable CAR then the

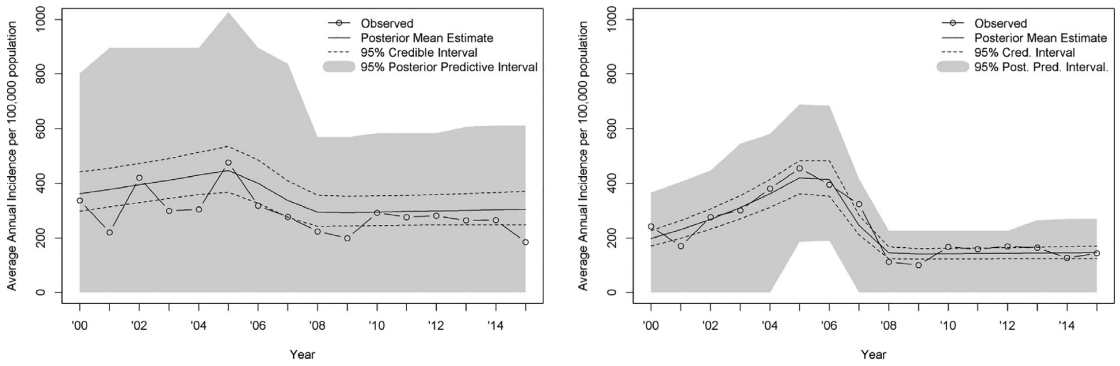


Fig. 6. Posterior mean of campylobacteriosis incidence estimates. The 95% credible interval is given in dashed lines and the 95% posterior predictive interval in grey.

Table 3
Posterior mean parameter estimates from the piecewise regression of the common CAR model, with their 95% credible intervals in parenthesis. α_p is the Poisson intensity at time zero while t^*_{1p} and t^*_{2p} are the estimated breakpoints. β_{0p} is the trend parameter for time interval 1. β_{1p} and β_{2p} are the additive changes in trend for time interval 2 and 3 respectively. β_{3p} is the deprivation score effect. Note, that all the effects are on a log-scale.

Parameter	*p=Rural*	*p=Urban*
α_p	−5.67 (−5.87, −5.46)	−6.38 (−6.53, −6.24)
β_{0p}	0.04 (0.02, 0.07)	0.15 (0.14,0.17)
β_{1p}	−0.21 (−0.30, −0.13)	−0.67 (−0.73, −0.62)
β_{2p}	0.18 (0.09,0.26)	0.53 (0.46,0.59)
t^*_{1p}	2005.2 (2004.7,2005.8)	2005.7 (2005.6,2005.9)
t^*_{2p}	2007.9 (2007.3,2008.6)	2008.1 (2007.9,2008.3)
β_{3p}	−0.03 (−0.07,0.01)	−0.006 (−0.03,0.02)

temporal model. In what follows, we will report the estimates of the best model only.

The posterior mean estimates and 95% credible intervals for the piecewise regression parameters are given in Table 3. For rural areas, β_{0p} shows that the campylobacteriosis incidence has increased by an average of approximately 4%, (2%, 7%) CI, per year in the first time interval (2000–2005). The estimated posterior mean for the second time interval ($\beta_{0p} + \beta_{1p}$) shows that the incidence then decreased by 17% per year (2006–2008) and then again increased by 0.67% per year in the third time interval (2009 onward).

For urban areas, β_{0p} shows than the average campylobacteriosis incidence increased by approximately 15% per year in time interval one (2000–2005). The posterior mean for ($\beta_{0p} + \beta_{1p}$) shows that incidence then decreased by 52% per year in time interval two (2006–2008). The incidence then increased by 0.62% per year in time interval three (2009 onward).

The deprivation index was found to have a very slight negative effect, which was not significantly different from zero, for both, rural and urban areas. The posterior probability estimate for the deprivation effect is $P(\beta_{3rural} < \beta_{3urban}) = 0.85$. As this value is close to 1, it is highly likely that the deprivation effect in rural areas is smaller than that in urban.

The posterior mean estimates with their 95% credible interval and 95% posterior predictive intervals are displayed in Fig. 6. The posterior mean campylobacteriosis incidence is also superimposed in the area specific time series plots and is given by the dark thick line see Fig. 7. Figs. 6 and 7 show that the log-linear piecewise regression capture the temporal dynamics well.

The observed annual average incidence of campylobacteriosis for 2000–2015, posterior mean estimates and their exceedance probabilities are displayed in Fig. 8. The posterior mean estimates are pulled towards local averages therefore producing smoother (shrunken) incidence estimates. Exceedance probabilities can be used to investigate possible areas with higher disease risk (Best et al., 2005; Lawson and Rotejanaprasert, 2014). CAUs with exceedance probabilities close to 1, indicate that the incidence will likely be above expectation and can therefore be considered as high risk areas. Column three of Fig. 8 show that high risk CAU's are present in rural areas (shown by SDHB). High risk areas are also present in Invercargill City, the Queenstown Lakes, and Dunedin City. As the SDHB has many small urban areas next to large rural areas, it can be difficult to ascertain if an urban area has a high exceedance probability. Therefore Fig. 9 marks the areas with high exceedance probabilities and hence can be considered high risk.

4. Discussion

The key findings of this study indicate that the temporal dynamics of campylobacteriosis incidence differed between rural and urban areas. Campylobacteriosis incidence increased between 2000 and 2005 and was higher for urban areas (15%) compared to rural areas (4%). The decline in campylobacteriosis incidence for the 2006–2008 period was greater for urban areas (−52%) compared to rural areas (−17%). This suggests that the 2006 regulatory changes impacted urban and rural areas differently. This is important as it further highlights the disparity of risk factors between urban and rural areas. Further investigations are needed to better understand factors that drive these differences. By doing so, action plans can then be implemented to further reduce campylobacteriosis incidence in rural areas.

The objective of this study was to evaluate difference in campylobacteriosis incidence between rural and urban areas. The analysis was implemented using geocoded reported incidence that provided address accuracy at CAU level. However, Fig. 5 illustrated that there were a substantial number of areas that had zero reported incidence for a given year. Despite this, CAU's were chosen as the boundaries as it was the only level of areal aggregation that differentiated between urban and rural areas.

The data was analysed using the Besag, York and Mollié (BYM) CAR model which is well established in the field of epidemiology. However there are other Bayesian spatial models that can be implemented. These include the multivariate normal with exponential correlation (EXP), the spatial mixture model by Green and Richardson (MIX), and Knorr-Held and Raßer's partition Model (KHR) (Green and Richardson, 2002; Held and Rafer, 2000;

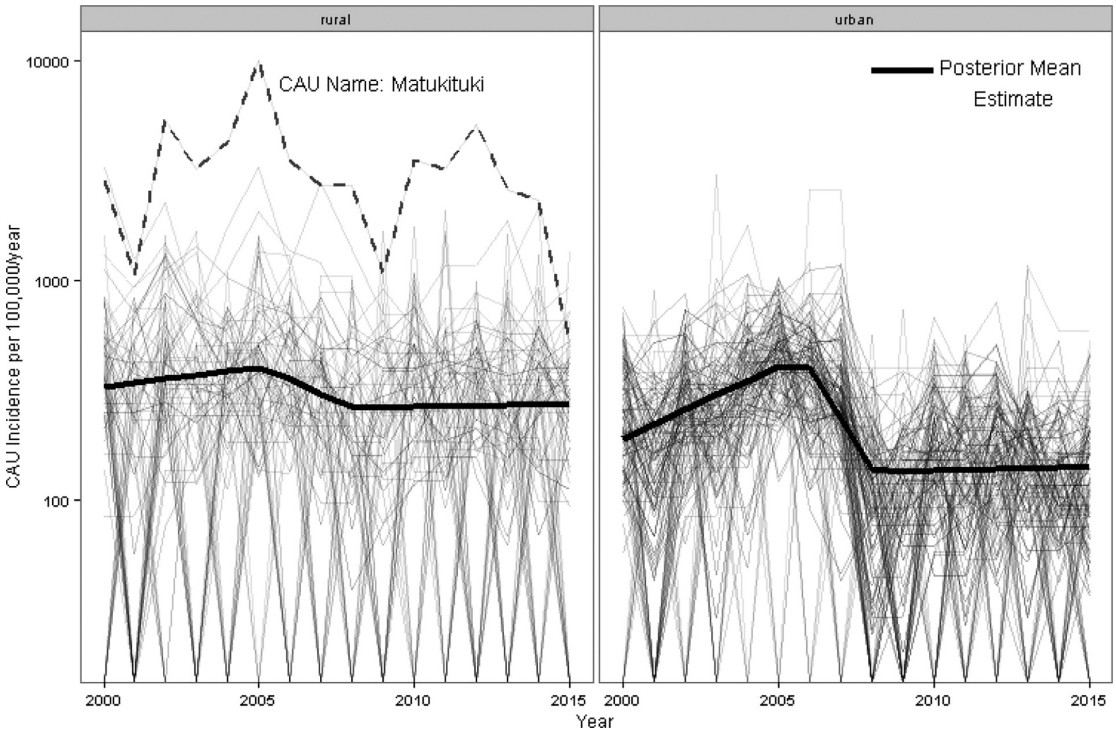


Fig. 7. Posterior mean estimates of campylobacteriosis incidence superimposed on the individual CAU time series plots.

Knorr-Held et al., 2000). In a comparative study conducted by Best et al., the authors found that the CAR performed well when modelling a single disease (Best et al., 2005). The authors also note that the EXP model over smoothed incidence and lead to poor inference compared to the CAR, KHR and MIX models. Furthermore they also state that the CAR and KHR model performed better at overall area risk classification but the MIX model produced less biased results for high risk areas. However unlike the CAR model, the MIX and KHR model assumes stationarity of the mean and variance of the spatial residual, which may be unrealistic in many applications. Further details of these comparisons can be found in the paper (Best et al., 2005).

It would have been of interest to inspect the spatial precision of the spatial effect of the temporal CAR model. By inspecting the spatial precision we could conclude if the incidence between areas changed over the study period. However as this model did not converge this was not possible.

Analysis was carried out to detect disease hot spots or clusters. In simpler terms these are areas which have an increased risk of campylobacteriosis incidence. To identify high risk areas, exceedance probabilities were used. High probabilities meant campylobacteriosis incidence for that area was likely to exceed expected rates, in turn suggesting that these areas face an increased disease risk. The plots in Fig. 8 illustrated the presence of high risk CAU's in rural areas. High risk areas were also present in Invercargill City, Queenstown Lakes, and Dunedin City. Fig. 9 also showed the locations of urban areas with high exceedance probabilities. These hot spots and clusters are of interest as they show areas that appear to have higher disease risk. These areas should be further examined to explain the higher incidence rates. By doing so, there is potential to uncover unknown risk factors and how they affect the aetiology of campylobacteriosis.

In this study, the results showed that deprivation was negatively associated with campylobacteriosis incidence and is consistent with previous studies (Gillespie et al., 2008; Nichols et al., 2012; Spencer et al., 2011b). The posterior probability estimate for the deprivation effect was $P(\beta_{3rural} < \beta_{3urban}) = 0.85$. As this value is close to 1, it is highly likely that the deprivation effect in rural areas is smaller than that in urban areas. This suggests that deprivation and urban/rural classification may affect notification rates. However the overall effect of deprivation for both urban and rural is inconclusive as zero is included in the credible intervals.

One of the major drawbacks of the study was the absence of area specific covariates such as occupation. Due to confidentiality restrictions this type of information was not available at CAU level and therefore could not be included in the model. Information on occupation may be relevant as there is evidence to suggest that agricultural workers have a higher probability of contracting campylobacteriosis (Gilpin et al., 2008; Savill et al., 2003; Spencer et al., 2011b). By using occupational information, the high risk CAU's in rural areas may be explained (see Fig. 8).

Ecological fallacy or bias occurs when it is assumed that the correlation at the aggregated level transfers to the individual level (Greenland and Robins, 1994). The effect of ecological bias can cause misleading results; in some cases the association between the phenomena and covariate(s) can disappear or reverse. Ecological fallacy is a known and common issue in disease mapping (Lokar et al., 2019; Richardson et al., 2003; Wakefield, 2007; Wang et al., 2017). Therefore when associations between covariates and a disease are found using the spatial CAR model, it is often followed by cohort-based or case-control based individual comparisons to investigate if these associations persist. In our study we did not find evidence to suggest that deprivation effected campylobacteriosis incidence. However, this outcome maybe changed if an individual case-control study was used instead.

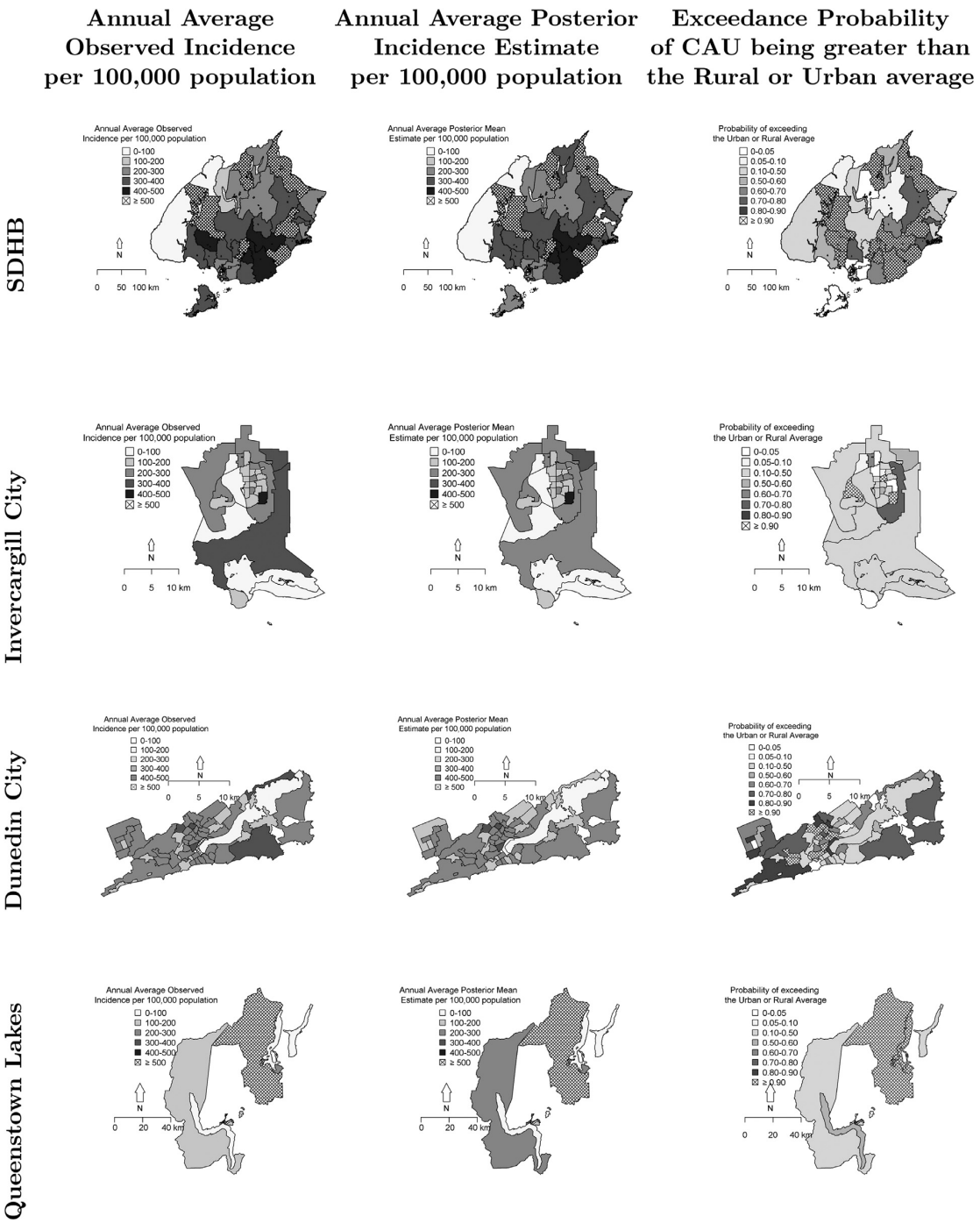


Fig. 8. Maps that compare observed incidence annual average, their posterior mean estimates and corresponding exceedance probabilities. The first column is the annual average observed incidence per 100,000 population for 2000–2015. The second column is the annual posterior mean and the third column is the probability that an area will exceed expected rates.

The work conducted here shows that it is important to consider the differing campylobacteriosis risk factors between urban and rural areas. Further research should be undertaken to better understand why rural areas still face a higher risk of campylobacteriosis despite the changes to the poultry industry. By investigating this further, regulatory authorities and policy makers can make more informed decisions and implement changes to reduce the risk to human health. In the future, we plan to extend the spatial analysis to the entire of New Zealand to investigate the differences between urban and rural areas and to uncover other high risk areas.

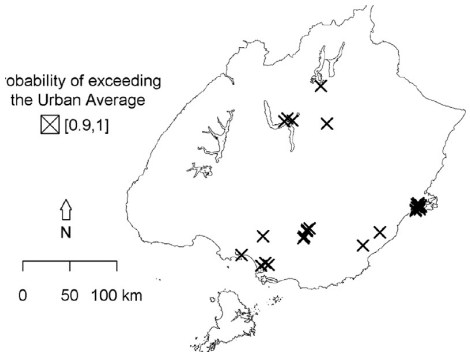


Fig. 9. Urban CAUs with high probability of exceeding the urban average for any time period.

Acknowledgements

Rodelyn Jaksons was jointly supported by the ESR (Institute of Environmental Science and Research Ltd.) Postgraduate scholarship and UC Connect Doctoral Scholarship (The University of Canterbury, New Zealand).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.sste.2019.100304.

References

Atkinson, J., Salmond, C., Crampton, P., 2014. NZDep2013 Index of Deprivation. Technical Report. University of Otago.

Baker, M.G., Sneyd, E., Wilson, N.A., 2007. Is the major increase in notified campylobacteriosis in New Zealand real? Epidemiol. Infect. 135, 163–170. doi:10.1017/S0950268806006583.

Besag, J., York, J., Mollié, A., 1991. Bayesian Image Restoration, with two applications in Spatial Statistics. Ann. Inst. Statist. Math 43 (1), 1–59.

Best, N., Richardson, S., Thomson, A., 2005. A comparison of Bayesian spatial models for disease mapping. Stat. Methods Med. Res. 14, 35–59. doi:10.1191/0962280205sm388oa.

Gillespie, I.A., O'Brien, S.J., Penman, C., Tompkins, D., Cowden, J., Humphrey, T.J., 2008. Demographic determinants for Campylobacter infection in England and Wales: implications for future epidemiological studies. Epidemiol. Infect. 136 (12), 1717–1725. doi:10.1017/S0950268808000319.

Gilpin, B.J., Scholes, P., Robson, B., Savill, M.G., 2008. The transmission of thermotolerant Campylobacter spp. to people living or working on dairy farms in New Zealand. Zoonoses Public Health 55 (7), 352–360. doi:10.1111/j.1863-2378.2008.01142.x.

Green, P.J., Richardson, S., 2002. Hidden Markov models and disease mapping. J. Am. Stat. Assoc. 97 (460), 1055–1070. doi:10.1198/016214502388618870.

Greenland, S., Robins, J., 1994. Invited commentary: ecologic studies - biases, misconceptions, and counterexamples. Am. J. Epidemiol. 139 (8), 747–760.

Held, L., Rafer, G., 2000. Bayesian Detection of Clusters and Discontinuities in Disease Maps. Biometrics 56 (1), 13–21.

Institute of Environmental Science and Research Limited, 2009. Notifiable and Other Diseases in New Zealand 2008 Annual Surveillance Report. Technical Report. Institute of Environmental Science and Research Limited, Porir.

Institute of Environmental Science and Research Limited, 2012. Enteric Disease: Manual for Public Health Surveillance in New Zealand ESR. Technical Report. Institute of Environmental Science and Research Limited, Porirua.

Knorr-Held, L., Rasser, G., Raß er, G., 2000. Bayesian detection of clusters and discontinuities in disease maps. Biometrics 56 (1), 13–21.

Lal, A., Lill, A.W.T., McIntyre, M., Hales, S., Baker, M.G., French, N.P., 2015. Environmental change and enteric zoonoses in New Zealand: A systematic review of the evidence. Austr. New Zealand J. Public Health 39 (1), 63–68. doi:10.1111/1753-6405.12274.

Lawson, A.B., Rotejanaprasert, C., 2014. Childhood brain cancer in Florida: a Bayesian clustering approach. Stat. Public Policy (October 2015) 00. doi:10.1080/2330443X.2014.970247.

Levesque, S., Fournier, E., Carrier, N., Frost, E., D. Arbeit, R., Michaud, S., 2013. Campylobacteriosis in urban versus rural areas: acase-case study integrated with molecular typing to validate risk factors and to attribute sources of infection. PLoS ONE 8 (12), 17–20. doi:10.1371/journal.pone.0083731.

Lokar, K., Zagar, T., Zadnik, V., 2019. Estimation of the ecological fallacy in the geographical analysis of the association of socio-economic deprivation and cancer incidence. Int. J. Environ. Res. Public Health 16, 296.

Lunn, D., Thomas, A., Best, N., D., S., 2000. Winbugs a bayesian modelling framework: concepts, structure, and extensibility.. J. Stat. Softw. 10, 325–337.

Martinez-Beneito, M.a., García-Donato, G., Salmerón, D., 2011. A Bayesian Joinpoint regression model with an unknown number of break-points. Ann. Appl. Stat. 5 (3), 2150–2168. doi:10.1214/11-AOAS471. arXiv: 1112.1526v1.

Ministry of Health, 2017. Notifiable diseases. https://www.health.govt.nz/our-work/diseases-and-conditions/notifiable-diseases.

Muggeo, V.M.R., 2003. Estimating regression models with unknown break-points. Stat. Med. 22 (19), 3055–3071. doi:10.1002/sim.1545.

Mullner, P., Spencer, S.E., Wilson, D.J., Jones, G., Noble, A.D., Midwinter, A.C., Collins-Emerson, J.M., Carter, P., Hathaway, S., French, N.P., 2009. Assigning the source of human campylobacteriosis in New Zealand: a comparative genetic and epidemiological approach. Infect. Genet. Evol. 9 (6), 1311–1319. doi:10.1016/j.meegid.2009.09.003.

New Zealand Law Resources, 1956. Health Act 1956.

New Zealand Law Resources, 2000. New Zealand public health and disability act 2000.

Nichols, G.L., Richardson, J.F., Sheppard, S.K., Lane, C., Sarran, C., 2012. Campylobacter epidemiology: a descriptive study reviewing 1 million cases in England and Wales between 1989 and 2011. BMJ Open 2, 1–13. doi:10.1136/bmjopen-2012-001179.

Pattis, I., Lopez, L., Cressey, P., Horn, B., Roos, R., 2017. Annual Report Concerning Foodborne Disease in New Zealand 2016, 2017: ESR Client Report FW17008, Christchurch, New Zealand. Technical Report.

Porirua: Institute of Environmental Science Research, 2011. EspiSurv User Guide. Technical Report. Institute of Environmental Science and Research, Porirua.

R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Richardson, S., Thomson, A., Best, N., Elliott, P., 2003. Interpreting posterior relative risk estimates in disease-mapping studies. Environ. Health Perspect. 112 (9), 1016–1025.

Salmond, C., Crampton, P., Atkinson, J., 2002. NZDep2006 Index of Deprivation User's Manual. Technical Report. University of Otago.

Salmond, C., Crampton, P., Atkinson, J., 2006. NZDep2001 Index of Deprivation User's Manual. Technical Report. University of Otago.

Savill, M., Hudson, A., Devane, M., Garrett, N., Gilpin, B., Ball, A., 2003. Elucidation of potential transmission routes of Campylobacter in New Zealand.. Water Sci. Technol. 47 (3), 33–38.

Sears, A., Baker, M.G., Wilson, N., Marshall, J., Muellner, P., Campbell, D.M., Lake, R.J., French, N.P., 2011. Marked campylobacteriosis decline after interventions aimed at poultry, New Zealand. Emerg. Infect. Dis. 17 (6), 18. doi:10.3201/eid1706.101272.

Soleimani, A., Hassanzadeh, J., Motlagh, A.G., Tabatabaee, H., Partovipour, E., Keshavarzi, S., Hossein, M., 2015. Spatial analysis of common gastrointestinal tract cancers in counties of Iran.. Asian Pacific J. Cancer Prevent. 16 (9), 4025–4029. doi:10.7314/APJCP.2015.16.9.4025.

Southern District Health Board,. About Southern DHB.

Spencer, S.E., Marshall, J., Pirie, R., Campbell, D., French, N.P., 2011. The detection of spatially localised outbreaks in campylobacteriosis notification data. Spatial Spatio-Temporal Epidemiol. 2 (3), 173–183. doi:10.1016/j.sste.2011.07.008.

Spencer, S.E.F., Marshall, J., Pirie, R., Campbell, D., Baker, M.G., French, N.P., 2011. The spatial and temporal determinants of campylobacteriosis notifications in New Zealand, 2001-2007. Epidemiol. Infect. 140, 1663–1677. doi:10.1017/S0950268811002159.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B 64 (4), 583–616. doi:10.1111/1467-9868.00353. arXiv: 1011.1669v3.

Statistics New Zealand, a. Census. Technical Report.

Statistics New Zealand, b. Classifications and related statistical standards. http://archive.stats.govt.nz/methods/classifications-and-standards/classificationrelated-stats-standards.aspx.

Statistics New Zealand, 2006a. Population mobility of urban / rural profile areas. http://archive.stats.govt.nz/browseforstats/population/Migration/internal-migration/mobility-urban-rural-areas.aspx.

Statistics New Zealand, 2006b. Urban/rural profile (experimental) classification categories. http://archive.stats.govt.nz/methods/classifications-and-standards/urbanrural-profile-experimental-class-categories.aspx.

Sturtz, S., Ligges, U., Gelman, A., 2005. R2winbugs: a package for running winbugs from r. J. Stat. Softw. 12 (3), 1–16.

Tiwari, R.C., Cronin, K.a., Davis, W., Feuer, E.J., Yu, B., Chib, S., 2005. Bayesian model selection for join point regression with application to age-adjusted cancer rates. J. R. Stat. Soc. Ser. C 54 (5), 919–939. doi:10.1111/j.1467-9876.2005.00518.x.

Wakefield, J., 2007. Disease mapping and spatial regression with count data. Biostatistics 8 (2), 158–183.

Waldhoer, T., Wald, M., Heinzl, H., 2008. Analysis of the spatial distribution of infant mortality by cause of death in Austria in 1984 to 2006.. Int. J. Health Geograph. 7, 21. doi:10.1186/1476-072X-7-21.

Wang, F., Wang, J., Gelfand, A., Li, F., 2017. Accommodating the ecological fallacy in disease mapping in the absence of individual exposures. Stat. Med. 36 (August), 4930–4942. doi:10.1002/sim.7494.

Zhuang, L., Cressie, N., 2012. Spatio-temporal modeling of sudden infant death syndrome data. Stat. Methodol. 9 (1-2), 117–143. doi:10.1016/j.stamet.2011.01.006.

Underreporting of disease risk

In epidemiology, researchers often want to estimate the number of people infected with the disease in a given period. Usually, the reported cases underestimate the actual burden. The data for many infectious diseases, such as gastroenteritis and influenza, are known to be underreported as the majority of those infected usually recover at home [160, 148, 80, 133]. For chronic conditions such as cancer and diabetes with a long period of latency, the recorded incidence, prevalence and mortality may also be underestimated. Underreporting of rare or chronic disease can be a result of misdiagnosis. For Type I diabetes, many people who are over 30, are often misdiagnosed as type II, as the early symptoms of both diseases are similar [124, 176, 8]. In rare diseases such as cancer, diagnosing cases is difficult as there are few screening tests and the ones available vary in accuracy [101, 10, 29]. Moreover, some individuals may also seek medical advice too late or lack symptoms until the disease has progressed [17, 33]. Other reasons for underreported cancer data can be because of administration issues. For example, in the USA, some studies have shown that cancer rates may be underestimated due to incomplete reporting of cases, or because of reporting delays [111, 11].

Underreported data are ubiquitous in all disciplines. In ecology, researchers are often interested in estimating the species abundance. However, the data may underestimate the true abundance as a result of not all areas being sampled or due to non-detection [155, 134]. In criminology, when assessing the number of crimes such as theft, assaults etc., conclusions are based only on what is reported, but the correct figures are known to be higher [56, 90, 106]. For some neighbourhoods and minority groups, underestimation of crime may be higher, as the victims may fear retribution from those responsible

if they were to report it[136, 79, 178].

In an economic study by Winkelmann in 1996, the researchers state that worker absenteeism caused by working conditions is problematic as it decreases output. It is argued that absenteeism decreases if staff are paid appropriately and have fair working conditions. To test this hypothesis, the authors use German socio-economic panel data. However, they note that if the employer reports the absenteeism, they may face recollection difficulties and so may underestimate the overall figure. On the other hand, if the data are employee-based, they may only report absenteeism linked to severe illness [197].

There are a variety of ways to account for the underreporting and to estimate the unobserved number of cases. For example, in ecology, it has been shown that the use of capture-recapture is useful in determining population size. Capture-recapture methods have also been used in epidemiology when two disease registries or data sets are available, and a proportion of participants are known to cross over [83, 147, 109, 26]. However, multiple data sources are often not available and therefore can not be used in many applications.

In ecology, presence-only data is a common set-up. For example, in citizen science data, the individual may report on seeing a kiwi bird in one location they visited. Still, nothing may be known about the other visited or unvisited locations. Therefore, it is unknown whether a kiwi was seen but not reported whether there were no kiwis present in the immediate area. The presence-only problem in ecology is similar to underreported data in epidemiology. For example, if we take the disease campylobacteriosis, we can never be sure whether a municipal area with zero reported cases truly did not have any campylobacteriosis cases, or whether there were unreported campylobacteriosis cases. Thus, methods used for presence-only data in ecology may be useful in dealing with underreported epidemiological data. Currently, the most well-known and popular method to deal with presence-only data in ecology is MaxEnt [49, 140, 138,

139]. MaxEnt estimates an areas habitat suitability for a given species using a suitable covariate and assumes that the presence locations are from a random sample. It evaluates a habitat suitability index by calculating the ratio of the conditional covariate density at the presence sites, to the marginal covariate density across the spatial domain [49, 140, 138, 139] The literature on MaxEnt is fraught with controversy [82, 156]. One of the criticisms is that MaxEnt assumes a random sample, which presence-only data may violate. Locations are usually visited because there is a higher chance of observing the species [193, 199]. Due to similarities with underreported epidemiological data, and presence-only data, MaxEnt has been used in epidemiological studies to uncover high-risk areas [5, 95, 119].

Given the population at risk N, the number of cases is typically modelled using a binomial distribution. When the true number of cases is unknown, we may instead assume that the observed number of cases Y has a binomial distribution:

$$Y \sim Bin(Z, p),$$

where, Z is the true number of cases, and p is the probability of detection/reporting. Here, both Z and p are unknown. In 1971 Draper and Guttman discussed how Bayesian statistics could be used to estimate the parameters in a binomial distribution when both the probability p and size Z was unknown. They recommend that the data should be used to estimate Z and to elicit prior knowledge on the probability p [45]. In 1987, Adrian Raftery proposed an empirical Bayes approach, which exploited the knowledge that the true number is at least what is observed, with his model allowing for interval estimation, prediction and point estimation [145].

The alternative to the binomial distribution, when N is large, and Np is small, is the Poisson distribution. If the data are underreported, the rate parameter can be adjusted for the missing observations. Winkelmann implemented this approach in 1996, to estimate the underreporting in worker absenteeism [197], by Moreno and Giron to

assess the overall burden of crime [122], and by Schmertmann and Gonzaga to correct for underestimated mortality schedules in Brazil [161]. More recently, Oliveira et al. developed the random censoring Poisson model and compared its performance against the model proposed by Moreno and Giron [128, 122]. Their model employed a two-step process, where, using a suitable covariate, they first calculated the probability that an area's count was censored. The second step was to estimate the true count for censored areas. This model is advantageous when a select number of areas are known to suffer in data quality, and a suitable covariate is available for estimating its censoring probability [128]. However, in cases where it is believed that there is censoring or under-reporting in all areas, the first step is skipped altogether.

In this study III, the observed counts Y, form only a subset of Z, the true number of cases. The random variable Z is an unobserved and unknown quantity and is treated as a latent variable. To account for the underreporting, we propose the use of Bayesian hierarchical models to estimate the latent variable Z. Also, we can infer features about the disease such as its underlying risk, and identify which areas are more likely to suffer from underreporting. To illustrate how the method works, we apply the model to the Pennsylvania Lung Cancer data set, where we assume a constant detection probability of $\phi = 0.9$. Sixteen different scenarios were simulated to see how the model would perform in different situations.

9.1 Data

Pennsylvania lung cancer data set

We apply the model to Pennsylvania Lung Cancer Data available from the **SpatialEpi** package in **R** [92, 143]. Pennsylvania is a state in the USA situated in the north east (see *Figure 9.1*).

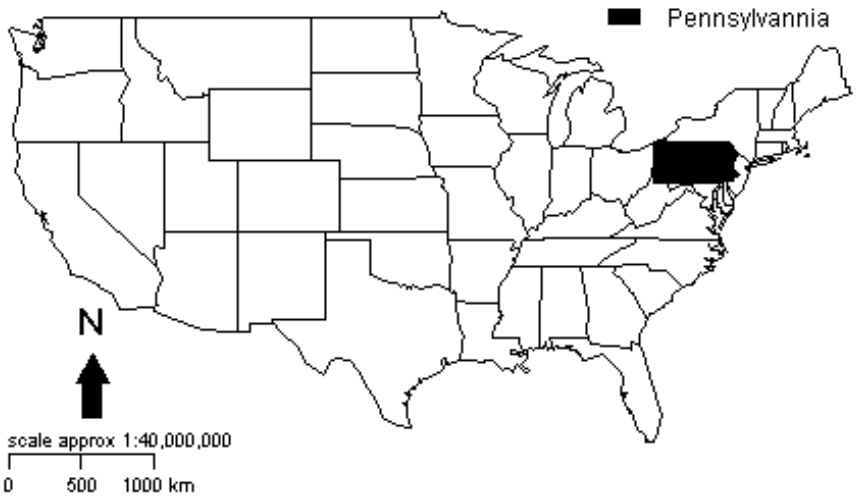


FIGURE 9.1. Location of Pennsylvania in the USA.

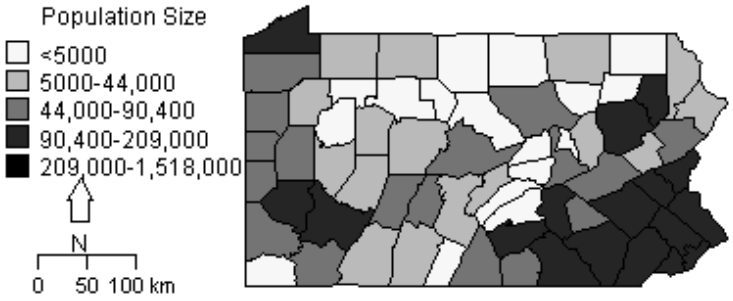
The data set consists of the lung cancer counts reported in 2002 at the county level with $n = 67$. For each county, the total population and the proportion of smokers has also been recorded. A table of summary statistics is displayed in *Table 9.1*.The county-level lung cancer counts and smoking proportions were obtained from the Pennsylvania Department of Health website [84], while the population counts were from the 2000 decennial census [30]. The data was stratified on race (white vs non-white), gender, and age (Under 40, 40-59, 60-69 and 70+). Although the demographics mentioned above are known to to be associated with lung cancer risk, we chose to aggregate the data to county-specific level, to illustrate how the model works [62, 13, 141, 144].

		Quantiles					
	Mean (sd)	Min	25%	50%	75%	Max	Total
reported cases	150 (240)	3	34	70	170	1,400	10,279
population	18,0000 (270,000)	4,900	44,000	90,000	210,000	1,500,000	12,281,054
smoking	0.24 (0.024)	0.18	0.23	0.23	0.26	0.28	0.24
counties	-	-	-	-	-	-	61

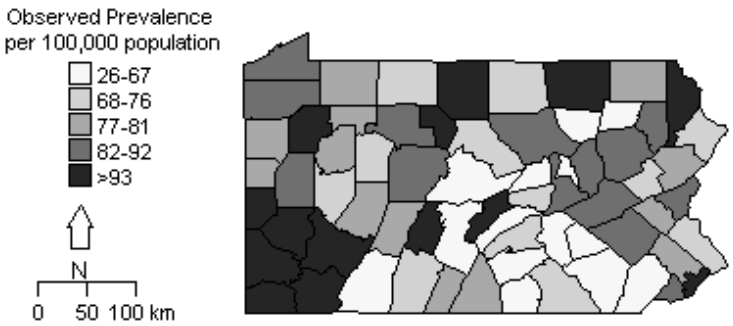
TABLE 9.1. Summary statistics of the Pennsylvania lung cancer data.

The population count of each county is shown in *Figure 9.2a* with highly populated counties situated in the south-east. The lung cancer incidence rates and smoking rates are plotted in *Figure 9.2b* and *Figure 9.2c* respectively. In *Figure 9.2b*, it is shown that there is a cluster of high observed prevalence in the south-west. *Figure 9.2c* shows that smoking rates are highest in the north-west, as well as the north-central counties. However, there is little variability in smoking rates between counties, with the lowest

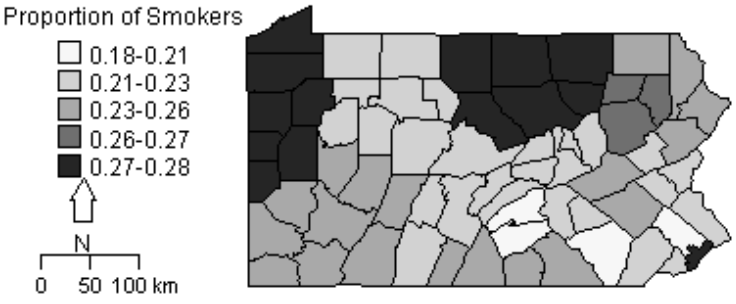
proportion of smokers $p = 0.18$, and the highest being $p = 0.27$, of the county population.



(A) County population



(B) Observed lung cancer prevalence, per 100,000 population



(C) Proportion of smokers

FIGURE 9.2. County based summary statistics.

9.2 Methodology

Let the observed number of cases Y be a subset of the true number of cases Z.

Let Z, follow a binomial distribution with population at risk N, and disease risk $\lambda(X)$.

$$Z \sim Binomial(N, \lambda(X)) \tag{9.1}$$

$\lambda(X)$ is a function of the smoking covariate and is expressed through an inverse-logit function or expit regression function with:

$$\lambda(X) = \frac{\exp\{X\beta + \varepsilon\}}{1 + \exp\{X\beta + \varepsilon\}}. \tag{9.2}$$

Where, X the vector of the county-specific smoking proportion, and β is a vector of the corresponding regression coefficients. The expit function is used to ensure the disease risk probability stays between zero and one. A spatial residual or spatial random effect ε can be included to account for spatial autocorrelation. In this study, it was assigned a CAR normal prior:

$$\varepsilon_j \sim N(\bar{\varepsilon}_{-j}, \tau m_j). \tag{9.3}$$

Here, the spatial precision is denoted by τ, and is given the prior $\tau \sim Gamma(0.1, 0.1)$, while m_j is the number of neighbours for area j [23]. Further details of the CAR prior can be found in *Chapter 3*. The regression coefficients are given non-informative normal priors with $\beta_i \sim N(0, 0.04)$, where the scale parameter is the precision.

Let the observed number of cases Y follow a binomial distribution with size Z, the true number of cases, and detection probability $\phi(X)$.

$$Y \sim Binomial(Z, \phi(X)), \tag{9.4}$$

where, $\phi(X)$ can also be a function of known covariates and can be expressed through an expit regression function with

$$\phi(X) = \frac{\exp\{X\alpha + \varepsilon\}}{1 + \exp\{X\alpha + \varepsilon\}}. \tag{9.5}$$

Here, X can be known covariates, and α is a vector of the corresponding regression coefficients. To account for spatial autocorrelation, correlated errors ε can also be added to the regression function, which is assigned an appropriate spatial prior. In this application, covariates are not available in estimation of detection ϕ. Thus, the regression for ϕ simplifies to an intercept only model, and an informative normal prior is placed on α with $\alpha \sim N(logit(0.9), 100)$.

By this formulation, the observed values Y are such that $Y \leq Z$. The resulting likelihood for the observed values Y follows a binomial distribution with population at risk N, and probability $\lambda(X) \cdot \phi(X)$.

$$f(Y|N, \phi(X), \lambda(X)) \sim Binomial(N, \phi(X) \cdot \lambda(X)) \tag{9.6}$$

The proof of the likelihood derivation can be found in the appendix $A.1$. The joint posterior distribution for the disease risk λ and detection probability ϕ is given by:

$$f(\lambda, \phi|Y, N) \propto f(Y|N, \lambda, \phi) f(\lambda) f(\phi). \tag{9.7}$$

The posterior for the true number of cases Z is obtained by

$$f(Z|Y, N) \quad \propto \int_0^1 \int_0^1 \binom{N}{Z} \lambda^Z (1-\lambda)^{N-Z} \binom{Z}{Y} \phi^Y (1-\phi)^{Z-Y} f(\lambda) f(\phi) d\phi d\lambda. \tag{9.8}$$

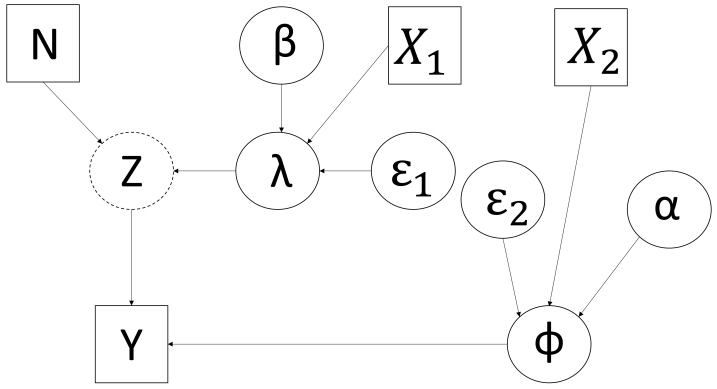


FIGURE 9.3. Graphical representation of the model.

Under this formulation, the intercepts in $\lambda(X)$ and $\phi(X)$ are unidentifiable, so informative priors are needed on at least one of the intercepts. In this study, an informative normal prior was placed on the detection probability ϕ, more specifically on the intercept α in the regression function.

The directed acyclic graph (DAG) of the proposed model is given in *Figure 9.3*. The dashed node indicates that Z is a latent variable.

Hierarchical models can induce large a posteriori correlations between the parameters. This results in the conditional variances of some variables being much smaller than the marginal variances. The sampler then exhibits a random walk type of behaviour, which explores the target distribution slowly [24]. The use of a non-centred parameterisation or Hamiltonian-Monte Carlo is recommended to reduce the correlation between the levels of hierarchy [24]. A non-centred parameterisation adds auxiliary variables to the location parameter of the prior distribution, which shifts the correlations from the latent parameters to the data, which allows the sampler to search the entire parameter space effectively [24].

Thus, when the trace plot of the regression parameters exhibited slow mixing, a non-centred parameterisation was implemented. The non-centred parameterisation yields

	Prior distribution for ϕ	Prior distributions for β_i	Prior distribution for ε_i
Non-spatial model	$\phi \sim N(logit(0.9, 100))$	$\beta_i \sim N(0, 0.04)$	$\varepsilon \equiv 0$
spatial CAR	$\phi \sim N(logit(0.9, 100))$	$\beta_i \sim N(0, 0.04)$	$\varepsilon_j \sim N(\overline{\varepsilon}_{-j}, \tau m_j)$ $\tau \sim Gamma(0.1, 0.1)$
spatial non-centred CAR	$\phi \sim N(logit(0.9, 100))$	$\beta_i \sim N(0 + \pi_i\psi_i, 0.04)$ $\pi_i \sim N(0, 0.05)$ $\psi_i \sim Gamma(1, 1)$	$\varepsilon_j \sim N(\overline{\varepsilon}_{-j}, \tau m_j)$ $\tau \sim Gamma(0.1, 0.1)$

TABLE 9.2. The prior distributions used in the model variations

$$\beta_i \sim N(\mu + \pi_i\psi_i, 0.04) \tag{9.9}$$

$$\pi_i \sim N(\mu_{\pi_i}, v) \tag{9.10}$$

$$\psi_i \sim Gamma(a, b) \tag{9.11}$$

$$\tag{9.12}$$

In this study, covariates are considered for the disease risk λ, and detection ϕ is treated as an intercept only model. Additionally, three variants of the model were fitted which were, 1) a non-spatial model which assumes $\varepsilon_i \equiv 0$, 2) a spatial model with a CAR component to take into account any spatial effects, and 3) and a CAR model which uses a non-centred parameterisation of the regression coefficients. In this study, detection does not have any uncertainty surrounding it, so additional dispersion parameters were not included. For clarity, the prior distributions for the model variants are tabulated in *Table 9.2*.

Spatial autocorrelation is assumed to be present if the spatial models indicated a better fit. Model selection was based on deviance information criterion (DIC), where models with smaller DIC values are preferred [169].

The model was fitted using **WinBUGS** through the **R2WinBUGS** package in **R** [105, 179, 143]. The model was run for 500,000 iterations, with a burn-in of 250,000 iterations,

and a thinning rate of five. Convergence was determined by visual assessments of trace plots and marginal posterior distributions.

9.3 Results

The posterior estimates for the model parameters are summarised in *Table 9.3*. The posterior mean estimates for the model parameters are similar between the models. However, the posterior standard deviation for β_1 increases, and the 95% credible interval becomes wider with increasing model complexity. For example, the posterior standard deviation for β_1 is higher for the non-centred CAR model compared to the others.

All three models agree that smoking is positively correlated with cancer prevalence, as the 95% credible intervals for β_1 do not include zero. The non-centred CAR model is the preferred model as it has the lowest DIC value, which also indicates that spatial autocorrelation is present. The resulting maps shall be based on that model.

Model	$\hat{\beta}_0$ (sd)	$\hat{\beta}_0$ CI	$\hat{\beta}_1$(sd)	$\hat{\beta}_1$ CI	$\hat{\phi}$(sd)	$\hat{\phi}$ CI	DIC
non-spatial	-6.981(0.014)	(-7.008,-6.953)	2.643(0.353)	(1.937,3.348)	0.900(0.009)	(0.881,0.916)	615.39
CAR	-7.022(0.018)	(-7.057,-6.985)	1.974(0.657)	(0.750,3.230)	0.900(0.009)	(0.881,0.916)	524.30
non-centred CAR	-7.024(0.018)	(-7.060,-6.986)	2.00(0.9044)	(0.189,3.809)	0.900(0.009)	(0.880,0.917)	489.44

TABLE 9.3. The Posterior Mean estimates of the regression parameters for incidence rate $\lambda(X)$ (β_0, β_1), detection rate ϕ, and the corresponding DIC of the fitted models.

Figure 9.4a depicts the posterior mean estimate for the true prevalence per 100,000 population. It predicts a cluster of high lung cancer risk in the south-west, as well as in the central north, and northeast. *Figure 9.4b* shows the difference between the predicted true prevalence, and the observed prevalence (both standardised per 100,000 population). The larger the difference between the two values, the higher the chance of underreporting. The figure shows that counties in the west, in particular, the north-west, have a higher risk of underestimating cancer prevalence.

Figure 9.4c shows the posterior mean disease risk estimate λ_i, where the disease risk is estimated to be the highest in the west, with a cluster in the east. *Figure 9.4d* shows the model residuals and depicts the unexplained noise after accounting for all relevant information. The darker areas show increased unexplained noise, which appears to be clustered in the south-east and the south-west.

Figure 10.4b are of the posterior exceedance probabilities, with probabilities close to one, identifying high-risk areas. The model estimates high disease risk areas are present in the south-west, which can also imply that there may be risk factors that are unaccounted for.

9.4 Simulated case studies

To get an indication of how the model will perform under different scenarios, we simulated 16 different data sets. The data sets refer to situations, where the disease is rare versus not rare, and when underreporting or detection is high versus when it is low. In this section, we will describe how the data sets were simulated, and report on the outcomes.

Simulation procedure

We simulated 16 different data sets that represent different scenarios. We use the state of Pennsylvania USA as our study area with the state split at the county level with $n = 67$.

The true number of cases is given by Z, which follows a binomial distribution with population at risk N, and disease risk $\lambda(X)$

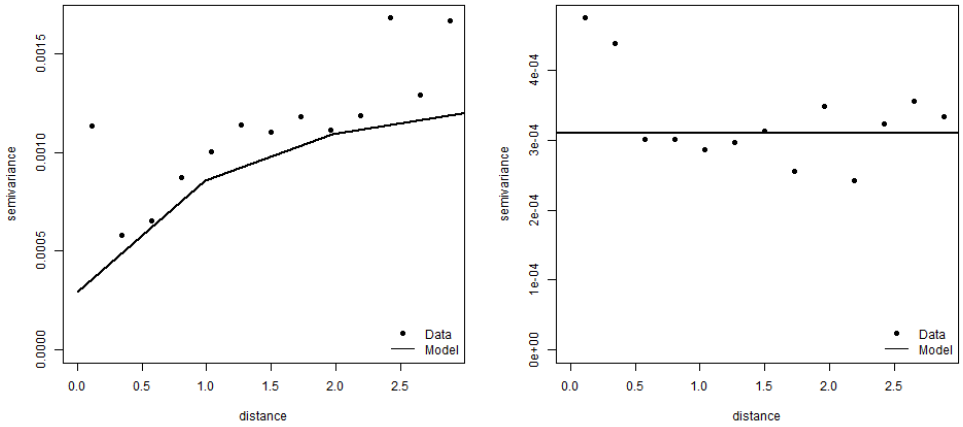
$$Z \sim Binomial(N, \lambda(X)). \tag{9.13}$$

The parameter $\lambda(X)$ is described as an expit regression function:

$$\lambda(X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \varepsilon_i)}{1 + \exp(\beta_0 + \beta_1 X_1 + \varepsilon_i)} \tag{9.14}$$

The covariate X_1, follows a normal distribution with $X_1 \sim N(0, 0.024^2)$. The variance of the simulated covariate was based on the variance of the smoking covariate in the Pennsylvania lung cancer data set, see *section 7.2*.

The parameter ε is an additional dispersion parameter, and takes the following values $\varepsilon = \{\varepsilon_1, \varepsilon_2\}$. The term ε_1 denotes spatial autocorrelation, and ε_2 represents a nugget-only effect. To generate the spatial autocorrelation, $\varepsilon = \varepsilon_1$, the centroid of each county was used to generate an isotropic random field of spatially correlated errors from an exponential semivariogram model with parameters $d = 2.50$, $nugget = 0.0003$, $sill = 0.003$, see *Figure 9.5a*. The distance d is the range parameter, which corresponds to the 65th percentile of the Euclidean distances between centroids. In the case where there is only a nugget-only effect, i.e., no spatial autocorrelation, $\varepsilon = \varepsilon_2$, a pure nugget semivariogram model was simulated, see *Figure 9.5b*.



(A) Variogram of the residuals in the spatial model

(B) Variogram of the residuals in the non-spatial model

FIGURE 9.5. The variograms of the model residuals for the simulated data. *Figure 9.5a* displays the residuals when spatial autocorrelation is present. *Figure 9.5b* shows the variogram when no spatial autocorrelation is present.

The observed number of cases Y is randomly sampled from a binomial distribution with size Z, the true number of cases, and with detection probability ϕ:

$$Y \sim Binomial(Z, \phi). \tag{9.15}$$

Various combinations of the model parameters are explored to see how the model performs in different scenarios, and are given in *Table 9.4*.

Detection ϕ_i	Spatial autocorrelation: ε_1 $logit(\lambda) = \beta_0 + \beta_1 X + \varepsilon$ $\varepsilon_j \sim N(\bar{\varepsilon}_{-j}, \tau m_j)$		No spatial autocorrelation : ε_2 $logit(\lambda) = \beta_0 + \beta_1 X + \varepsilon$ $\varepsilon_j \equiv 0$	
	Parameter	Coefficient	Parameter	Coefficient
$\phi_i = 0.7$ $logit(\phi) = 0.7$	β_0	-6.7	β_0	-6.7
	β_1	0	β_1	0
	β_0	-6.7	β_0	-6.7
	β_1	1	β_1	1
	β_0	-2	β_0	-2
	β_1	0	β_1	0
	β_0	-2	β_0	-2
	β_1	1	β_1	1
$\phi_i = 0.9$ $logit(\phi) = 0.9$	β_0	-6.7	β_0	-6.7
	β_1	0	β_1	0
	β_0	-6.7	β_0	-6.7
	β_1	1	β_1	1
	β_0	-2	β_0	-2
	β_1	0	β_1	0
	β_0	-2	β_0	-2
	β_1	1	β_1	1

TABLE 9.4. Combinations of the model parameters to show case the different scenarios.

The parameters β_0 describes the baseline prevalence of a disease, with $\beta_0 = -6.7$ (prevalence=123 per 100,000), represents rare diseases such as cancer. When $\beta_0 = -2$ (prevalence=3533 per 100000), it represents more common diseases such as endometriosis in women [85]. The parameter for β_1 indicates whether the covariate is associated with disease prevalence, with $\beta_1 = 1$ stating that they are correlated, and $\beta_1 = 0$,

assumes no association. The detection probability ϕ describes the severity of underreporting, which takes different values with $\phi = \{0.7, 0.9\}$. In this simulation study, detection does not have any uncertainty surrounding it, so additional dispersion parameters were not included in its regression equation.

Model fitting procedure

In this section, the scale parameters for the normal priors is the precision. The parameters β_0 and β_1 are assigned non-informative normal priors, $\beta_i \sim N(0, 0.04)$. When $\phi = 0.9$, α was assigned a normal prior, with $\alpha \sim N(logit(0.9), 100)$. When $\phi = 0.7$, the prior for α was given a normal distribution, with $\alpha \sim N(logit(0.7), 100)$. Placing an informative prior was necessary for identifiability and convergence of the model. For the spatial process, the spatial random effect was assigned a CAR prior.

A non-centred parameterisation of the regression coefficients β_0 and β_1 in the CAR model was also fitted to achieve faster convergence, as described in *section 9.3*. For clarity, the prior distributions for the different models are shown in *Table 9.5*

The models were fitted in **WinBUGS**, via the **R2WinBUGS package** in **R** [105, 179, 143]. Five hundred thousand iterations were run with 250,000 used as burn-in. A thinning rate of 50 was applied to reduce the autocorrelation between iterations; with 5000 iterations used for model inference. Model convergence was visually assessed from trace plots and marginal posterior densities.

Results of simulated cases

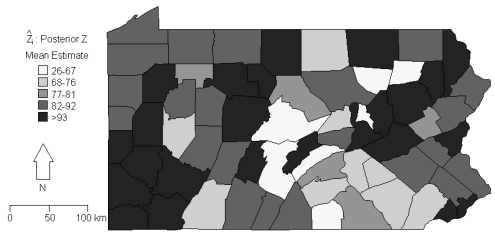
The posterior estimates for the model parameters of the different simulated scenarios are given in *Table 9.6*.

	Detection ϕ	Regression coeffiecients $\lambda(X)$	Dispersion ε
Non-spatial	$\phi = 0.9$ $\alpha \sim N(logit(0.9), 100)$	$\beta_i \sim N(0, 0.04)$	$\varepsilon_j \equiv 0$
Spatial CAR	$\phi = 0.9$ $\alpha \sim N(logit(0.9), 100)$	$\beta_i \sim N(0, 0.04)$	$\varepsilon_j \sim N(\bar{\varepsilon}_{-j}, \tau m_j)$ $\tau \sim Gamma(0.1, 0.1)$
Spatial: non-centred CAR	$\phi = 0.9$ $\alpha \sim N(logit(0.9), 100)$	$\beta_i \sim N(0 + \pi_i\psi_i, 0.04)$ $\pi_i \sim N(0, 0.05)$ $\psi_i \sim Gamma(1, 1)$	$\varepsilon_j \sim N(\bar{\varepsilon}_{-j}, \tau m_j)$ $\tau \sim Gamma(0.1, 0.1)$
Non-spatial	$\phi = 0.7$ $\alpha \sim N(logit(0.7), 100)$	$\beta_i \sim N(0, 0.04)$	$\varepsilon_j \equiv 0$
Spatial CAR	$\phi = 0.7$ $\alpha \sim N(logit(0.7), 100)$	$\beta_i \sim N(0, 0.04)$	$\varepsilon_j \sim N(\bar{\varepsilon}_{-j}, \tau m_j)$ $\tau \sim Gamma(0.1, 0.1)$
Spatial: non-centred CAR	$\phi = 0.7$ $\alpha \sim N(logit(0.7), 100)$	$\beta_i \sim N(0 + \pi_i\psi_i, 0.04)$ $\pi_i \sim N(0, 0.05)$ $\psi_i \sim Gamma(1, 1)$	$\varepsilon_j \sim N(\bar{\varepsilon}_{-j}, \tau m_j)$ $\tau \sim Gamma(0.1, 0.1)$

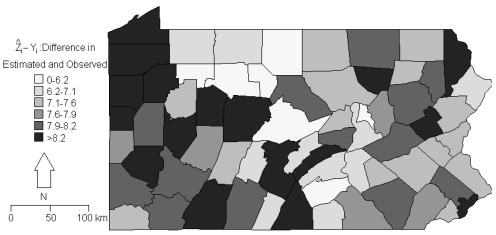
TABLE 9.5. Prior distributions of the model variants

The table shows that the posterior standard deviation for each parameter increases as detection decreases from $\phi = 0.9$ to $\phi = 0.7$. In general, the models provided good posterior estimates for all scenarios, with the 95% credible intervals providing reasonable values of the regression coefficients. The 95% credible intervals for each parameter are shown in *Figure 9.6*.

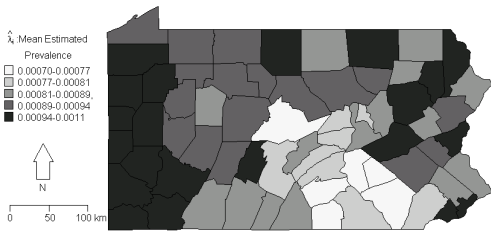
The posterior estimates for the model parameters of the non-centred parameterisation of the CAR model is tabulated in *Table 9.7*. It gives similar results as the original model, however for some parameters, the posterior standard deviation increases.



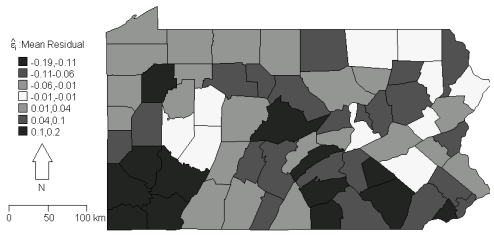
(A) Estimated posterior mean $\hat{Z}$: True Prevalence



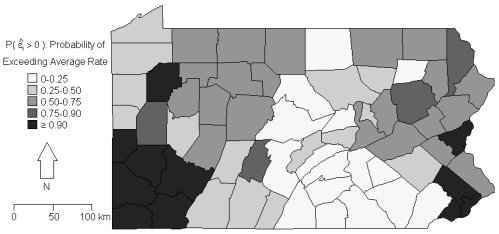
(B) Difference in True Estimated Prevalence and Observed Prevalence



(C) Estimated Posterior Prevalence Rate: $\hat{\lambda}_i$



(D) Model Residual: ε_i



(E) Probability of Exceeding Average Prevalence Rate: $P(\varepsilon_i > 0)$

FIGURE 9.4. Maps of the posterior mean estimates of the model parameters

	$\phi = 0.9$, $\beta_0 = -6.7$						$\phi = 0.7$ $\beta_0 = -6.7$					
Other Parameters	$\hat{\beta}_0$ (sd)	$\hat{\beta}_0$ CI	$\hat{\beta}_1$ (sd)	$\hat{\beta}_1$ CI	$\hat{\phi}$ (sd)	$\hat{\phi}$ CI	$\hat{\beta}_0$ (sd)	$\hat{\beta}_0$ CI	$\hat{\beta}_1$ (sd)	$\hat{\beta}_1$ CI	$\hat{\phi}$ (sd)	$\hat{\phi}$ CI
$\beta_1 = 2.5$, $\varepsilon = \varepsilon_1$	**-6.630 (0.015)**	**(-6.66,-6.599)**	2.92 (0.464)	(1.494,3.379)	0.900 (0.009)	(0.881,0.916)	-6.637 (0.033)	(-6.700, -6.570)	2.311 (0.495)	(1.381,3.305)	0.700 (0.021)	(0.658, 0.742)
$\beta_1 = 2.5$, $\varepsilon = \varepsilon_2$	-6.719 (0.016)	(-6.749, -6.688)	2.392 (0.486)	(1.494,3.379)	0.900 (0.009)	(0.881,0.916)	-6.713 (0.033)	(-6.777,-6.646)	2.079 (0.809)	(1.122, 3.109)	0.700 (0.021)	(0.658, 0.742)
$\beta_1 = 0$, $\varepsilon = \varepsilon_1$	**-6.666 (0.015)**	**(-6.696, -6.636)**	**-1.206 (0.464)**	**(-2.087, -0.263)**	0.900 (0.009)	(0.881,0.917)	**-6.630 (0.033)**	**(-6.694, -6.564)**	0.235 (0.508)	(-0.712, 1.246)	0.700 (0.021)	(0.658, 0.741)
$\beta_1 = 0$, $\varepsilon = \varepsilon_2$	-6.702 (0.016)	(-6.732, -6.670)	0.210 (0.477)	(-0.691, 1.176)	0.900 (0.009)	(0.881, 0.917)	-6.692 (0.033)	(-6.754, -6.626)	-0.421 (0.509)	(-1.375, 0.592)	0.700 (0.021)	(0.657, 0.741)
	$\phi = 0.9$, $\beta_0 = -2$						$\phi = 0.7$ $\beta_0 = -2$					
Other Parameters	$\hat{\beta}_0$ (sd)	$\hat{\beta}_0$ CI	$\hat{\beta}_1$ (sd)	$\hat{\beta}_1$ CI	$\hat{\phi}$ (sd)	$\hat{\phi}$ CI	$\hat{\beta}_0$ (sd)	$\hat{\beta}_0$ CI	$\hat{\beta}_1$ (sd)	$\hat{\beta}_1$ CI	$\hat{\phi}$ (sd)	$\hat{\phi}$ CI
$\beta_1 = 2.5$, $\varepsilon = \varepsilon_1$	**-1.949 (0.014)**	**(-1.975, -1.923)**	2.571 (0.178)	(2.218, 2.937)	0.900 (0.011)	(0.879, 0.920)	-1.959 (0.033)	(-2.020, -1.891)	2.489 (0.188)	(2.219, 2.875)	0.704 (0.021)	(0.663, 0.742)
$\beta_1 = 2.5$, $\varepsilon = \varepsilon_2$	-1.200 (0.015)	(-2.026, -1.970)	2.540 (0.167)	(2.202, 2.870)	0.898 (0.012)	(0.876, 0.921)	-1.996 (0.038)	(-2.063, -1.915)	2.472 (0.170)	(2.140, 2.809)	0.699 (0.023)	(0.651, 0.741)
$\beta_1 = 0$, $\varepsilon = \varepsilon_1$	**-1.954 (0.014)**	**(-1.979, -1.926)**	0.026 (0.176)	(-0.327, 0.370)	0.901 (0.011)	(0.879, 0.921)	-1.954 (0.032)	(-2.010, -1.877)	0.004 (0.185)	(-0.368, 0.365)	0.701 (0.020)	(0.655, 0.736)
$\beta_1 = 0$, $\varepsilon = \varepsilon_2$	-1.200 (0.014)	(-2.022,-1.965)	-0.173 (0.174)	(-0.519, 0.173)	0.899 (0.011)	(0.874, 0.920)	-1.989 (0.041)	(-2.058, -1.909)	-0.056 (0.180)	(-0.426, 0.290)	0.693 (0.025)	(0.646, 0.737)

TABLE 9.6. The results of the different simulation studies. In this table the mean and standard deviation of each parameter is given, along with their corresponding 95% Credible Intervals.

	$\phi = 0.9$, $\beta_0 = -6.7$						$\phi = 0.7$ $\beta_0 = -6.7$					
Other Parameters	$\hat{\beta}_0$ (sd)	$\hat{\beta}_0$ CI	$\hat{\beta}_1$ (sd)	$\hat{\beta}_1$ CI	$\hat{\phi}$ (sd)	$\hat{\phi}$ CI	$\hat{\beta}_0$ (sd)	$\hat{\beta}_0$ CI	$\hat{\beta}_1$ (sd)	$\hat{\beta}_1$ CI	$\hat{\phi}$ (sd)	$\hat{\phi}$ CI
$\beta_1 = 2.5$, $\varepsilon = \varepsilon_1$	**-6.631 (0.016)**	**(-6.661,-6.600)**	2.701 (0.564)	(1.622,3.833)	0.900 (0.009)	(0.880,0.916)	-6.637 (0.033)	(-6.699, -6.570)	2.161 (0.583)	(1.002,3.302)	0.700 (0.021)	(0.657, 0.740)
$\beta_1 = 2.5$, $\varepsilon = \varepsilon_2$	-6.720 (0.016)	(-6.751, -6.688)	2.039 (0.556)	(0.998,3.179)	0.900 (0.009)	(0.880,0.918)	-6.715 (0.033)	(-6.776,-6.646)	1.945 (0.588)	(0.777, 3.115)	0.701 (0.021)	(0.656, 0.740)
$\beta_1 = 0$, $\varepsilon = \varepsilon_1$	**-6.666 (0.016)**	**(-6.698, -6.636)**	**-0.983 (0.475)**	**(-1.938, -0.76)**	0.900 (0.009)	(0.881,0.917)	**-6.631 (0.033)**	**(-6.694, -6.564)**	0.257 (0.531)	(-0.772, 1.306)	0.701 (0.021)	(0.657, 0.742)
$\beta_1 = 0$, $\varepsilon = \varepsilon_2$	-6.702 (0.016)	(-6.733, -6.670)	0.200 (0.480)	(-0.749, 1.139)	0.900 (0.009)	(0.881, 0.917)	-6.692 (0.033)	(-6.752, -6.623)	-0.327 (0.541)	(-1.422, 0.737)	0.700 (0.021)	(0.656, 0.740)
	$\phi = 0.9$, $\beta_0 = -2$						$\phi = 0.7$ $\beta_0 = -2$					
Other Parameters	$\hat{\beta}_0$ (sd)	$\hat{\beta}_0$ CI	$\hat{\beta}_1$ (sd)	$\hat{\beta}_1$ CI	$\hat{\phi}$ (sd)	$\hat{\phi}$ CI	$\hat{\beta}_0$ (sd)	$\hat{\beta}_0$ CI	$\hat{\beta}_1$ (sd)	$\hat{\beta}_1$ CI	$\hat{\phi}$ (sd)	$\hat{\phi}$ CI
$\beta_1 = 2.5$, $\varepsilon = \varepsilon_1$	**-1.949 (0.014)**	**(-1.974, -1.921)**	2.534 (0.191)	(2.147, 2.892)	0.900 (0.011)	(0.878, 0.919)	-1.961 (0.044)	(-2.037, -1.867)	2.431 (0.206)	(2.030, 2.836)	0.705 (0.027)	(0.649, 0.753)
$\beta_1 = 2.5$, $\varepsilon = \varepsilon_2$	-1.998 (0.014)	(-2.023, -1.969)	2.499 (0.174)	(2.147, 2.840)	0.898 (0.011)	(0.876, 0.918)	-1.998 (0.040)	(-2.078, -1.917)	2.434 (0.183)	(2.084, 2.802)	0.700 (0.025)	(0.652, 0.751)
$\beta_1 = 0$, $\varepsilon = \varepsilon_1$	**-1.952 (0.014)**	**(-1.977, -1.921)**	0.026 (0.185)	(-0.347, 0.379)	0.899 (0.011)	(0.875, 0.920)	-1.947 (0.044)	(-2.022, -1.848)	0.009 (0.192)	(-0.382, 0.370)	0.697 (0.026)	(0.639, 0.744)
$\beta_1 = 0$, $\varepsilon = \varepsilon_2$	-1.997 (0.014)	(-2.022,-1.969)	-0.172 (0.174)	(-0.514, 0.170)	0.899 (0.011)	(0.877, 0.918)	-1.997 (0.040)	(-2.076, -1.913)	-0.061 (0.183)	(-0.424, 0.292)	0.698 (0.025)	(0.649, 0.748)

TABLE 9.7. The results of the different simulation studies from the non centred parametisations of the CAR model. In this table the mean and standard deviation of each parameter is given, along with their corresponding 95% credible intervals.

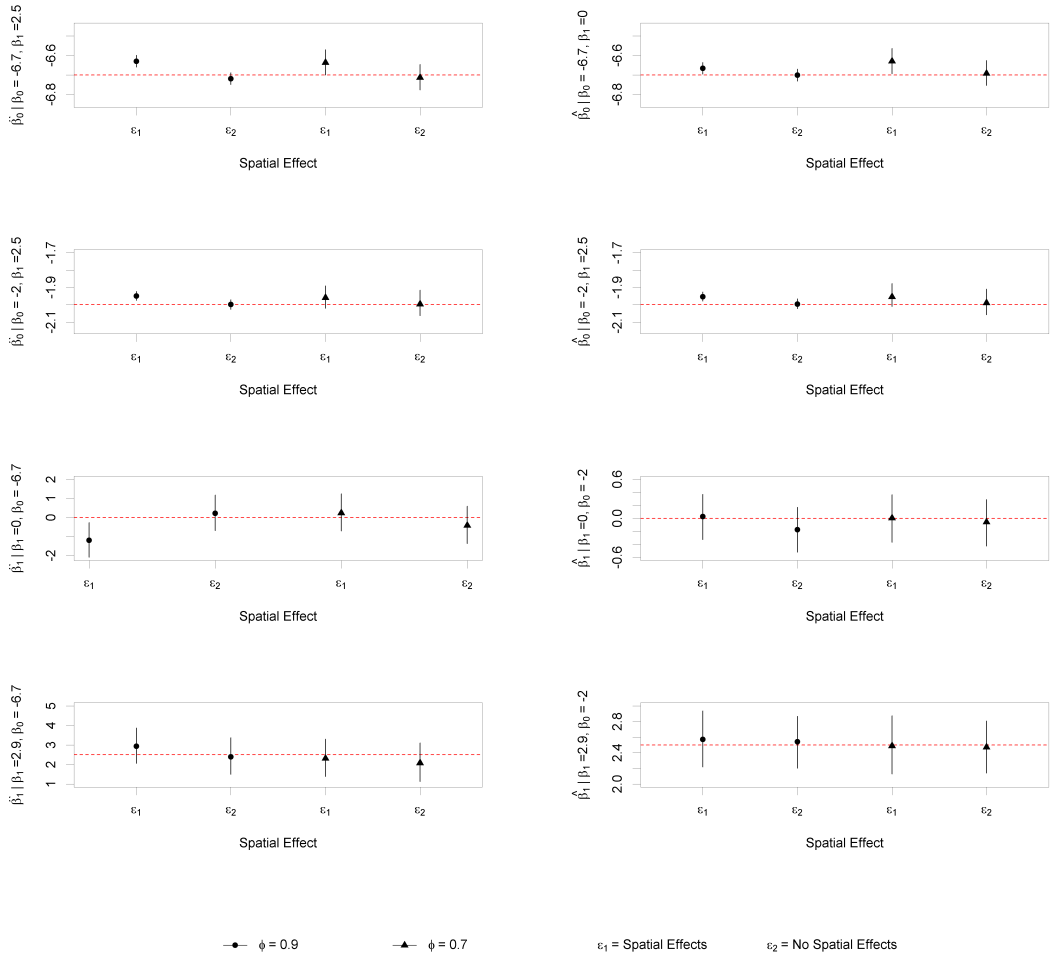


FIGURE 9.6. The 95% credible intervals of the parameters under different conditions, the red horizontal dash line gives the true value of the regression coefficient.

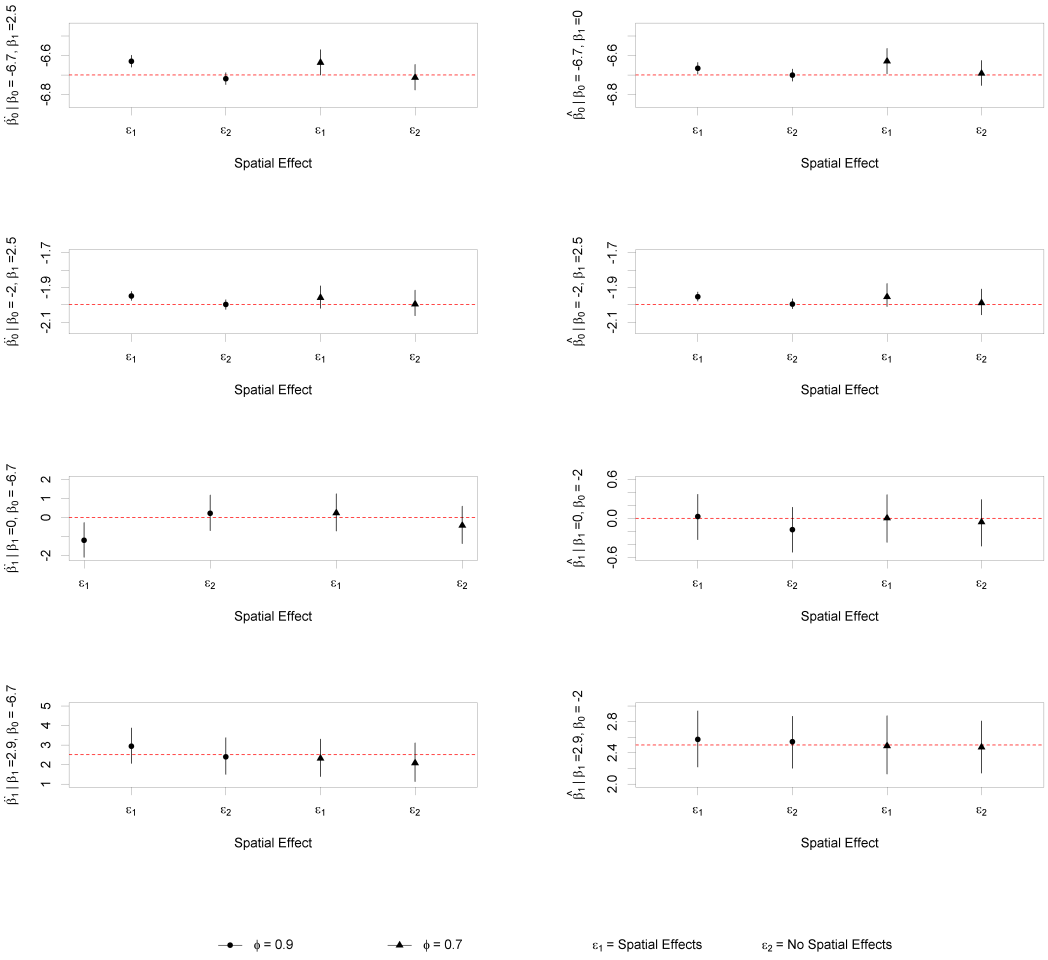


FIGURE 9.7. The 95% credible intervals of the parameters under different conditions for the non centred parametisation of the CAR model, the red horizontal dash line gives the true value of the regression coefficient.

9.5 Discussion

The objective of this study was to demonstrate how the model could be used for underreported count data. It is a promising method for estimating the unknown number of cases, as well as uncovering areas of increased underreporting.

The model comes with some caveats. First, the population at risk must be known. Moreover, the model needs informative priors on the baseline disease risk or detection probability. In many cases, such information may not be available. If a known binary

variable is available, the data can be split into sub-population, which could help address the issue of non-identifiability of some parameters [146, 78]. Another possibility is the use of capture recapture methods to estimate the detection probability, which can then be used to inform the prior distribution of detection. However, this is only possible if two disease registers are available, with some of the patients known to cross over.

The model was initially developed to estimate the prevalence of campylobacteriosis in New Zealand. However, it was not implemented, as no data was available on additional risk factors such as the proportion of the population engaging in agricultural work.

Instead, model suitability was tested on a stylised application to the Pennsylvania lung cancer data set, and 16 different simulated scenarios. The simulated case studies demonstrated how the models are capable of providing reasonable estimates for the model parameters. However, each situation was based on one simulated data set, and further simulation studies are required to get a better impression of model performance in general. In this thesis, additional simulation studies were not implemented due to time constraints; however, given more time and resources, this work can be extended.

9.6 Appendix

Proof of Likelihood

Let Z be the true number of cases which follows a binomial distribution with known population at risk N, and probability $\lambda(X)$. Let the observed number of cases Y, follow a binomial distribution with true number of cases Z and probability $\phi(X)$, such that the observed values Y are a subset of Z. The resulting likelihood, is the sum of a

product of two binomial likelihoods and is obtained by:

$$
\begin{aligned}
f(Y|N,\lambda(X),\phi(X)) &= \sum_{Z} f(Y,Z|N,\lambda(X),\phi(X)) \\
&= \sum_{Z} f(Y|Z,N,\theta(X),\phi(X)) f(Z|N,\theta(X),\phi(X)) \\
&= \sum_{Z} \binom{N}{Z} \lambda^{Z}(1-\lambda)^{N-Z} \binom{Z}{Y} \phi^{Y}(1-\phi)^{Z-Y} \\
&= \sum_{Z=Y}^{N} \binom{N}{Z} \lambda^{Z}(1-\lambda)^{N-Z} \binom{Z}{Y} \phi^{Y}(1-\phi)^{Z-Y} \\
&= \sum_{Z=Y}^{N} (\lambda\phi)^{Y} \lambda^{Z-Y} \binom{N}{Z} \binom{Z}{Y} (1-\phi)^{Z-Y}(1-\lambda)^{N-Z} \\
&= \sum_{Z=Y}^{N} (\lambda\phi)^{Y} \lambda^{Z-Y} \frac{N!}{Z!(N-Z)!} \frac{Z!}{Y!(Z-Y)!} (1-\phi)^{Z-Y}(1-\lambda)^{N-Z} \\
&= \sum_{Z=Y}^{N} (\lambda\phi)^{Y} \lambda^{Z-Y} \frac{N!}{Z!(N-Z)!} \frac{Z!}{Y!(Z-Y)!} \frac{(N-Y)!}{(N-Y)!} (1-\phi)^{Z-Y}(1-\lambda)^{N-Z} \\
&= \sum_{Z=Y}^{N} (\lambda\phi)^{Y} \lambda^{Z-Y} \frac{N!}{Y!(N-Y)!} \frac{(N-Y)!}{(N-Z)!(Z-Y)!} (1-\phi)^{Z-Y}(1-\lambda)^{N-Z} \\
&= \binom{N}{Y} (\lambda\phi)^{Y} \sum_{Z=Y}^{N} \binom{N-Y}{Z-Y} \lambda^{Z-Y}(1-\phi)^{Z-Y}(1-\lambda)^{N-Z} \\
&= \binom{N}{Y} (\lambda\phi)^{Y} \sum_{Z-Y=0}^{N-Y} \binom{N-Y}{Z-Y} [\lambda(1-\phi)]^{Z-Y}(1-\lambda)^{(N-Y)-(Z-Y)} \\
&= \binom{N}{Y} (\lambda\phi)^{Y} [\lambda(1-\phi)+(1-\lambda)]^{N-Y} \\
&= \binom{N}{Y} (\lambda\phi)^{Y} (1-\phi\lambda)^{N-Y}
\end{aligned}
$$

where $Y \leq Z \leq N$. The observed number of cases Y, is reduced to a binomial proportion with $Y \sim (N, \lambda \cdot \phi)$. For simplicity sake, the notation of (X) is eventually dropped in the proof.

Reducing the risk from western corn rootworm (*Diabrotica virgifera virgifera*)

10.1 Introduction

In Europe and North America, the western corn rootworm (WCR) beetle (*Diabrotica virgifera virgifera*) is a major agricultural pest. It is known to cause massive yield loss to Zea mays maize crops [112]. In places of prolonged infestation, the number of beetles observed has increased at an alarming rate. It is particularly problematic in Austria, where much maize is grown, and with little or no natural predators, the destruction continues unhindered.

It is essential to understand the emergence dynamics of established WCR beetle populations, to make effective pest management decisions. The WCR beetle is known to spend its egg stages in the soil, emerging in the late spring or early summer period [112, 28, 166]. Previous studies show that environmental factors such as temperature and precipitation influence the emergence cycle. For example, warmer temperatures increase the observed abundance, and WCR beetle emergence can last until the first frost [188], which in Austria usually occurs at the beginning of November. On the other hand, past research have reported that increased precipitation, and colder temperatures in the winter increase WCR beetle mortality [54, 187, 86, 22]. As is the case for many insects, the emergence dynamics of the WCR beetle can be described with sufficient accuracy by a parametric curve, such as, for example, the Gompertz curve [164].

In this study, we chose to use the Gompertz curve to model the observed emergence dynamics of the WCR beetle. The Gompertz curve was first proposed by Benjamin Gompertz in 1825 to describe the law of human mortality [72]. It is a sigmoidal curve which describes growth as being the slowest at the beginning and end of a period. It is usually characterized by three parameters; an asymptote, a relative starting value, and a growth rate coefficient which affects the slope. Since its introduction, the Gompertz model has been applied to many population biology studies [32, 3, 94, 174].

In this study, the asymptote parameter α is a proxy for the carrying capacity of the WCR beetle. The growth rate coefficient γ is an indicator of the emergence rate, with lower values indicating a protracted period of beetle emergence.

Only the traps with at least one non-zero observation were included into the model. Our model, therefore, only represents the regions with already established WCR beetle population. We incorporate the above structure into a Bayesian hierarchical modelling framework. We use Markov Chain Monte Carlo (MCMC) methods for parameter estimation and posterior inference. We apply the model to the WCR beetle capture data for Austria obtained in 2014 and investigate the effect of climate covariates such as temperature and precipitation on the WCR beetle population dynamics. We finally discuss handling of the missing data, as well as the accounting for spatial autocorrelation.

10.2 Data

The data set consists of records of WCR beetle captures across 204 maize-growing locations in Austria, as shown in *Figure 10.1* [2]. A trap was placed at a location where the WCR beetle had been previously observed or was expected to be seen [149]. The traps were laid at the beginning of the maize growing season (usually in the beginning of June) until harvest (usually the beginning of October), thus giving a monitoring period of 19 weeks. The traps operated by releasing pheromones which attracted male beetles and yellow sticky tapes captured the WCR beetle [93]. Each week, the number of beetles caught were counted and the traps emptied. The data were collected during

the monitoring period of 2014, with only observed cases included, i.e. traps that captured at least one WCR beetle over the entire summer of 2014.

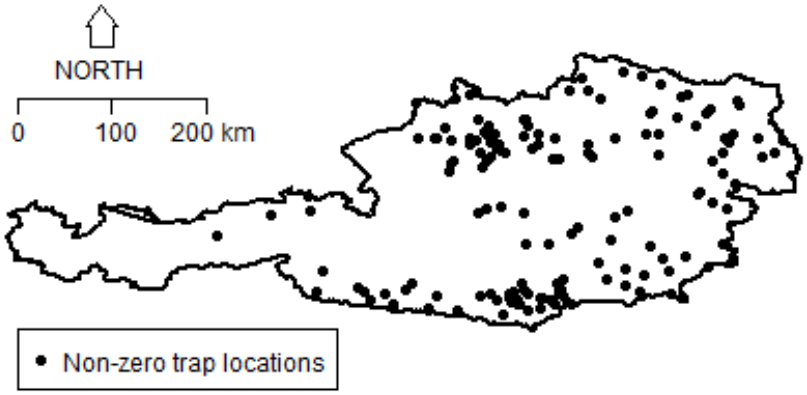


FIGURE 10.1. The locations of the placed traps where at least one WCR beetle was caught.

Of the total of 204 traps, monitoring WCR beetles in 2014, a total of 160 traps (78.43%) had at least one non-zero record and were thus included the analysis. For the remaining 160*19=3040 records, 713, (23.45%) were recorded as either blanks or zeroes. Although in most cases, this occurred in either the beginning or the end of the season. There were also 52 suspicious blanks/zeroes occurrences in the middle of the season. Because we were modelling cumulative emergence, omitting a missing data point would mean omitting the records for the remainder of the season for the entire trap and thus losing ten traps or 6.25% of the data. Instead, after a consultation with two domain experts, we set up the following scheme:

- Any blank or missing observation until the first numeric entry was coded as zero ($n = 606$, or 19.93%)
- Any blank or zero records which occurred between two non-zero entries, at least one of which was greater than or equal to 10, were recoded as missing. ($n = 40$, or 1.31%). Otherwise, they were coded as zeroes.
- Any blank or missing records which occurred between two zero entries were coded as zeroes ($n = 2$, or 0.07%).

- Any blank or missing observations that were between at least one non zero entry were coded as missing ($n = 10$, or 0.33%).

- The trap observations were only included in the analysis until the last numeric entry, which excluded ($n = 652$, or 21.4%) observations

Hence, if an observation was deemed missing, it was estimated as a parameter in the model. After implementing the scheme above, there were 2327 non-missing observations in the data set.

Climatic variables were available for each trap location based on the nearest weather station [67]. In this work the following variables are considered: the average winter temperature during winter (1st January - end of March), the average spring temperature (1st April - end of June), the cumulative precipitation (mm) during winter (1st January - end of March), the percentage of maize share in cropland (maize), and the total WCR beetle trap count averaged over all the traps within a 40 km of the location in 2013 ($\bar{y}_{2013}$). To detrend the spatial surface, we also included the coordinates, x and y. As they were originally given in longitude and latitude, we projected them to the World Geodetic System 1984 (WGS84), using the **sp** package in **R** [135, 143].

10.3 Methodology

Model

Let y_{it} denote the WCR beetle count observed in week t for trap i, and assume it to follow a Poisson distribution with parameter μ_{it}

$$y_{it}|\mu_{it}, \sim Poisson(\mu_{it}) \tag{10.1}$$

The intensity parameter μ_{it} represents the rate of emergence for a given period. Instead of allowing it to depend purely on time t, a phenological variable of growing degree days (GDD) is used [200, 25, 7, 6]. Warmer temperatures are required for insect development.

GDD reflect the heat accumulation and are defined as an integral of warmth above the threshold temperature after a given start date:

$$GDD = \int (T(t) - T_{base})dt.$$

The above integral can be approximated by

$$GDD = \max\left(\frac{T_{max} - T_{min}}{2} - T_{base}, 0\right). \tag{10.2}$$

Where T_{min} is the minimum daily temperature, T_{max} is the maximum daily temperature, and T_{base} is a set base temperature. In this study, the base temperature was set at 10°C, as this is the minimum temperature required for beetle maturation [112]. The starting date was the beginning of April, which marks the start of the growing season.

The Gompertz function is defined as

$$f(z_t) = \alpha \exp(-\beta \exp(-\gamma z_t)). \tag{10.3}$$

Where, α is the upper asymptote, β is a relative starting value, γ is a growth rate coefficient which affects the slope, and z_t are the cumulative growing degree days. In this study, one can consider the asymptote as the carrying capacity of the WCR beetle population. Moreover, lower values of β suggests an earlier first emergence, while lower values of γ indicate a more extended emergence period. To investigate if there is an association between climatic variables and the emergence dynamics, the Gompertz curve parameters, α and γ, are treated as linear functions of weather-related covariates. In this framework, a spatial residual can be added in either α and γ if there is evidence to do so.

To reflect the nature of the emergence dynamics and to preserve shape, the parameters of the model are restricted to positive values such that $\alpha > 0$, $\beta > 0$, and $\gamma > 0$. The time at inflection or period of highest growth can be obtained as

$$T^* = \frac{\log(\beta)}{\gamma}. \tag{10.4}$$

The Gompertz function describes cumulative emergence. Thus, to describe the incremental emergence rate, the derivative of the Gompertz function can be used instead. Consequently, as the data consists of weekly counts, the rate function μ_{it} is better described by the derivative of the Gompertz function:

$$\log(\mu_{it}) = \log(\alpha_i) + \log(\gamma_i) + \log(\beta_i) + \gamma_i z_{it} - \beta_i \exp(-\gamma z_{it}). \tag{10.5}$$

The parameters α_i, β_i and γ_i are trap specific such that:

$$\log(\alpha_i) \sim N(\mu_{\alpha_i}, \tau_\alpha) \tag{10.6}$$

$$\log(\gamma_i) \sim N(\mu_{\gamma_i}, \tau_\gamma) \tag{10.7}$$

$$\beta_i \sim \exp(1). \tag{10.8}$$

Here, τ_α and τ_γ are the precision parameters of the prior distributions for α, and γ respectively. Moreover, the means of the distributions μ_{α_i}, and μ_{γ_i} can be expressed as functions of known covariates:

$$\mu_{\alpha_i} = a_0 + \mathbf{w}^T X_{\alpha_i} \tag{10.9}$$

$$\mu_{\gamma_i} = g_0 + u^T X_{\gamma_i} + e_i \tag{10.10}$$

Where a_0 is the intercept and $\mathbf{w}$, is a vector of the regression coefficients, and X_α are the location-specific covariates. The order of the predictors used in the regression of μ_α were average winter temperature, precipitation, maize share, $\bar{y}_{2013}$, and the quadratic

trend in the location coordinates based on x, y, x^2, y^2, and xy. The parameter g_0 is the intercept for the regression μ_γ, and u is the corresponding regression coefficient. Average spring temperature was used as the predictor for μ_γ. The parameter e_i is a location-specific residual, which accounts for spatial autocorrelation. We assume that the spatial residuals have a multivariate normal distribution:

$$\mathbf{e} \sim MVN(0, \Sigma) \tag{10.11}$$

where $\mathbf{\Sigma}$ is a variance-covariance matrix, described by a variogram model [16, 198, 43]. The variogram is defined as the variance of the difference between variable values at two locations i and j and is often assumed to depend only on the distance d_{ij} between them. Here. we chose to use the powered exponential variogram model:

$$\Sigma_{ij} = \exp\{(\phi d_{ij})^\kappa\}. \tag{10.12}$$

Where ϕ is the rate of decline of correlation with distance between points d_{ij}, and κ is the spatial smoothing parameter.

As the Gompertz curve is a non-linear function, there are identifiability issues with the parameters α, and γ. Therefore, informative priors were placed on the intercepts a_0, and g_0. They are assigned normal priors; $a_0 \sim (5.2, 10)$, which is the average number of WCR beetles caught for the entire region. The prior for the growth rate intercept was $g_0 \sim N(-0.31, 100)$. The prior for g_0 was decided by first fitting a model first used a diffuse prior for all parameters, and did not include a spatial random effect. The posterior mean estimate for g_0 under this model was used to inform the prior for the spatial model.

The rest of the regression coefficients were given non-informative normal priors $N(0, 0.001)$. The precision parameter for τ_α was assigned prior distribution $\tau_\alpha \sim gamma(0.001, 0.001)$, while τ_γ was given prior $\tau_\gamma \sim gamma(0.01, 0.01)$. A tighter prior was placed on τ_γ,

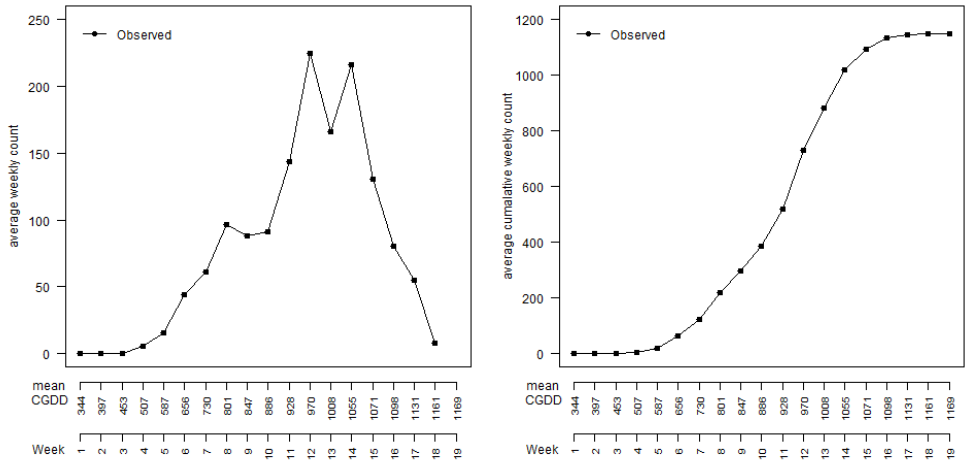
because there was less variability in the growth rate γ compared to the asymptote α, when examined empirically.

To determine the where to place the spatial residual, we used the residuals under the first model to fit variograms models. Based on this, we found no evidence for spatial autocorrelation in the emergence rate γ, but not for the carrying capacity α. The parameter ϕ was given an informative uniform prior $\phi \sim U((0.3, 0.75)$. The bounds of ϕ were chosen by fitting a parametric exponential variogram model to the growth rate residuals under the initial non-spatial model. The spatial precision τ was given a $\tau \sim gamma(0.01, 0.01)$ prior. The parameter κ was fixed at 1, which simplified to an exponential variogram model.

The model was fitted using **WinBUGS** through the **R2WinBUGS** package in **R** [105, 179, 143]. The model was run for 20,000 iterations, with a burn-in of 10,000 iterations, and a thinning rate of five. Convergence was determined by visual assessments of trace plots and marginal posterior densities.

10.4 Results

The raw observed mean weekly counts and mean cumulative counts for the entire study region are shown in *Figure 10.2a* and *Figure 10.2b* respectively. *Figure 10.2a* show that the highest number of trapped beetles took place between week 12 to week 15.



(A) Average weekly count

(B) Average weekly cumulative count

FIGURE 10.2. The observed weekly count (top left) and cumulative weekly count (top right).

To examine if spatial autocorrelation was present in the parameters α and γ; a non-spatial model was fitted first. The posterior mean residuals of the parameters α and γ were obtained, and sample semivariograms constructed. The residuals were defined as $\varepsilon_\alpha = \alpha_i - \mu_{\alpha_i}$ and $\varepsilon_\gamma = \gamma_i - \mu_{\gamma_i}$.

A sample semivariogram was constructed from every 20th iteration, and an exponential variogram model fitted. This procedure indicated that autocorrelation was present in γ; hence, an additional dispersion parameter, by way of the powered exponential function, was added to the regression of γ.

The parameter α, which represents the carrying capacity, was found to be positively correlated with temperature and precipitation. A 1°C increase in winter temperature was associated with an average expected carrying capacity increase of 19.24%, while a 1mm rise in cumulative precipitation was associated with an average expected carrying capacity decrease of 0.1%. However, the 95% credible interval for both coefficients included zero; thereby, a zero effect cannot be excluded. Maize share (maize) was also found to be positively correlated with α, with an increase in one percent of maize share

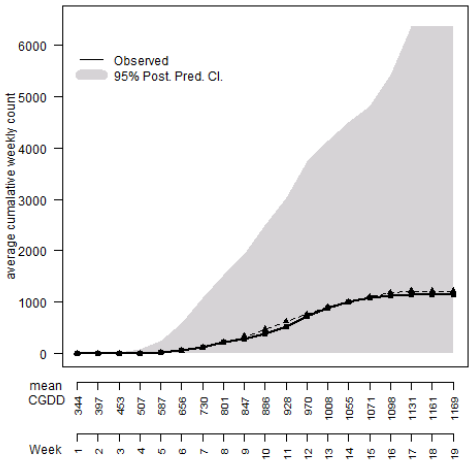
Parameter	Variable	Posterior Mean Estimate	Posterior Standard Deviation	95% Credible Interval
μ_α	intercept	5.150	0.227	(4.720,5.580)
	winter temperature	0.176	0.178	(-0.159,0.517)
	precipitation	-0.001	0.002	(-0.005,0.002)
	maize	0.040	0.01	(0.021,0.060)
	$\bar{y}_{2013}$	-0.003	0.002	(-0.006,0.0006)
	x	1.930	0.202	(1.550,2.320)
	y	-1.270	0.518	(-2.290,-0.278)
	xy	0.337	0.237	(-0.125,0.800)
	x^2	-0.308	0.084	(-0.463,-0.141)
	y^2	-0.497	0.348	(-1.180,0.172)
	τ_α	0.675	0.087	(0.518,0.854)
μ_γ	intercept	-0.272	0.071	(-0.418,-0.141)
	spring temperature	-0.294	0.052	(-0.402,-0.192)
	τ_γ	6.210	1.230	(4.250,8.960)
sill	τ	4.670	2.890	(1.440,12.700)

TABLE 10.1. The posterior estimates of the predictors on a log scale.

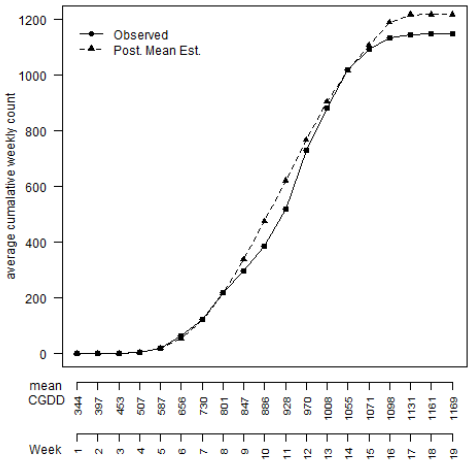
being associated with increasing the expected carrying capacity by 4.08%. The previous year's abundance ($\bar{y}_{2013}$) was on average negatively correlated with the expected rate, with the associated mean posterior decrease in carrying capacity of 0.30%. However, a zero effect is possible as reflected by the 95% credible interval.

The emergence rate coefficient γ was found to be negatively correlated with the average spring temperature, with a 1°C increase associated with an average 25.47% decrease in the expected growth rate. Therefore, increases in temperature translate into slower growth, implying a longer time to reach the asymptote and thus protracted beetle emergence in warmer springs. The posterior probability of the negative effect of average spring temperature on growth is $P = 1$. The time of inflection is defined by $T^* = \log(\beta)/\gamma$. Therefore, as warmer spring temperature give rise to smaller γ values, the time of inflection is expected to occur at higher cumulative GDD and the peak emergence can thus occur at a later date in the growing season.

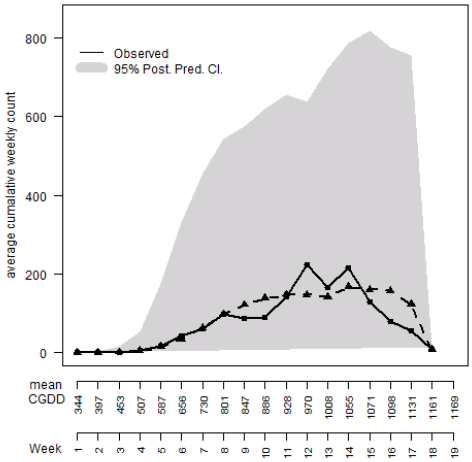
The model fits are shown in *Figure 10.3*. This figure shows that the 95% credible interval encapsulates the observed dynamics, and describes the data well.



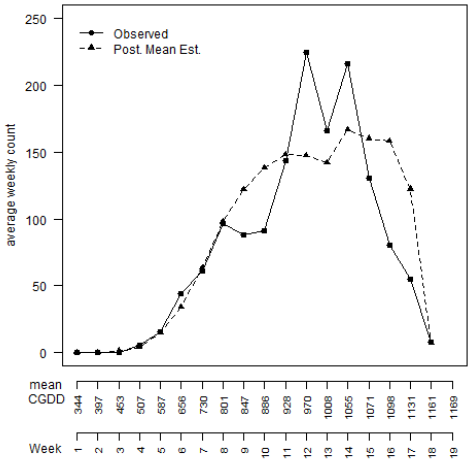
(A) 95% posterior predictive interval for mean cumulative counts



(B) Posterior mean estimate for the mean cumulative count and the raw mean cumulative count



(C) 95% posterior predictive interval for mean weekly count

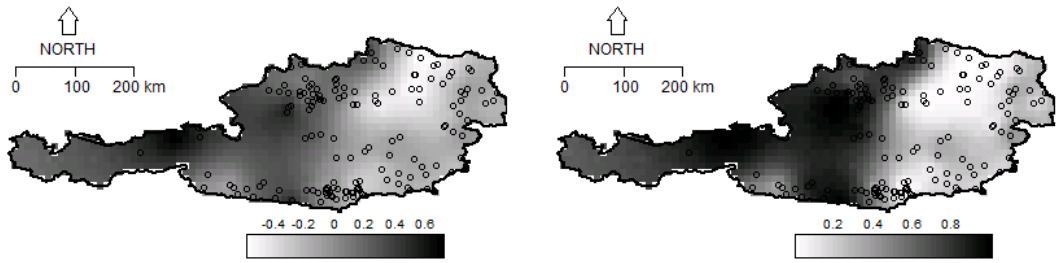


(D) Posterior mean estimate for meekly count and raw weekly mean count

FIGURE 10.3. Model Fit.

The interpolated surfaces for γ are in given in *Figure 10.4*. *Figure 10.4a* are the posterior predicted means of the growth rate spatial residual. *Figure 10.4b* depicts the probability of ε_γ exceeding zero. These plots show that the growth rate is above average in the west, and below average in the east. Therefore western areas will reach

the carrying capacity sooner, or have a shorter new emergence period. Whereas, in the east, there will be protracted emergence.



(A) Post. predicted means of ε_γ

(B) P(ε_γ >0)

FIGURE 10.4. Interpolation of γ residuals.

10.5 Discussion

The objective of this study was to better understand the emergence dynamics of established populations of the WCR beetle. Furthermore, it was of interest how climatic variables such as temperature and precipitation affected the observed dynamics. The results of the study showed that the carrying capacity was most affected by temperature.

Due to climate change, the average temperatures around the world have risen, and with this comes new challenges to agriculture. The WCR beetle population carrying capacity was found to be positively correlated with winter temperature. Therefore warmer winters are likely to result in larger WCR beetle populations and more severe infestation.

The rate of beetle emergence was found to be negatively correlated with spring temperatures, which means that increased temperatures protract the WCR beetle emergence period. For growers, this is problematic as it means their maize crops are subject to more extended periods of predation from the WCR beetle. Therefore, it may be beneficial to grow maize varieties that mature earlier. Alternatively, maize can be planted

earlier in the season, so that the rootstock is well established to withstand better damage caused by the WCR beetle.

Higher spring temperatures were also found to be associated with peak emergence to occurring later in the growing season. This information is vital for insecticide spraying, as it is usually timed according to peak emergence so that it is most effective in reducing population size [66, 103]. The adoption of these strategies is beneficial to growers as they have the best hopes of reducing yield loss.

The model suggested evidence of spatial autocorrelation for the emergence rate. Therefore, traps placed closer together were more alike in the emergence rate compared to traps further apart. Additionally, the interpolated surfaces of the emergence rate indicated that eastern areas would experience protracted new WCR beetle emergence.

Due to the nonlinearity of the Gompertz function, there are non-identifiability issues between the parameters, and informative priors are required. Our priors were constructed after consultation with domain experts. Because of non-identifiability issues, it is also impossible to place spatially autocorrelated residuals in sub-models for all the parameters. We chose to model the population emergence rate as a spatially autocorrelated phenomenon.

Missing data were an important issue in this data set. The primary problem was the absence of protocol for the data recording. The blanks and zeroes were used interchangeably for missing observations and actual zero counts, respectively. We came up with a reasonable re-coding scheme, but other choices are possible. Moreover, the choice of sampling locations did not follow any design. A better planned study may provide more insights into the phenomenon.

Finally, the analysis was based on the observed WCR beetle counts for traps with at least one non-zero count during the 2014 growing season. The results thus apply to

well-established pest populations only. It is vital to pay more attention to emerging populations as well to prevent further spread. In the future, we plan to analyse the longer term data from 2004-2015 to see not only how the WCR beetle emergence dynamics changes both spatially and temporally but also to get more insights into the dynamics of the population spread, continuing the work of Falkner et al. [54].

Bibliography

[1] "2003/766/EC: Commission Decision of 24 October 2003 on emergency measures to prevent the spread within the Community of Diabrotica virgifera Le Conte (notified under document number C(2003) 3880)". In: *Official Journal of the European Union* (2003).

[2] AGES. *Maiswurzelbohrer: Verbreitung. Maiswurzelbohrer Verbreit.* 2008. URL: https://www.ages.at/themen/schaderreger/maiswurzelbohrer/verbreitung/.

[3] J. J Ahn, C. Y Yang, and C Jung. "Model of *Grapholita molesta* spring emergence in pear orchards on statistical inforinform criteria". In: *Journal of Asia-Pacific Entomology* 15.4 (2012), pp. 589–593.

[4] H. Akaike. "A New Look at the Statistical Model Identification". In: *Automatic Control, IEEE Transactions on* 19.6 (1974), pp. 716–723. ISSN: 15582523. DOI: 10.1109/TAC.1974.1100705. URL: http://ieeexplore.ieee.org/search/srchabstract.jsp?tp=\&arnumber=1100705\&queryText=((Authors:akaike))\&openedRefinements=*\&sortType=desc_Publication+Year\&ranges=1974_1974_Publication_Year\&matchBoolean=true\&rowsPerPage=50\&searchField=Search+All.

[5] M. A. Alkhamis and K. VanderWaal. "Spatial and Temporal Epidemiology of Lumpy Skin Disease in the Middle East, 2012â€"2015". In: *Frontiers in Veterinary Science* 3.March (2016), pp. 1–12. DOI: 10.3389/fvets.2016.00019.

[6] W Anderson, R Smith, and J McWilliam. "A systems approach to the adaptation of sunflower to new environments II. Effects of temperature and radiation on growth and yield". In: *Field Crops Research* 1 (1978), pp. 153–163.

[7] J Angus et al. "Phasic development in field crops II. Thermal and photoperiodic responses of spring wheat". In: *Field Crops Research* 4 (1981), pp. 269–283.

[8] S Appel et al. "Latent autoimmune diabetes of adulthood (LADA): An often misdiagnosed type of diabetes mellitus". In: *Journal of the American Academy of Nurse Practitioners* 21.3 (2009), pp. 156–159.

[9] P. Aragón, A. Baselga, and J. M. Lobo. "Global estimation of invasion risk zones for the western corn rootworm Diabrotica virgifera virgifera: integrating distribution models and physiological thresholds to assess climatic favourability". In: *Journal of Applied Ecology* 47.5 (2010), pp. 1026–1035. URL: https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2664.2010.01847.x.

[10] M. Arbyn et al. "Pooled analysis of the accuracy of five cervical cancer screening tests assessed in eleven studies in Africa and India". In: *International Journal of Cancer* 123.1 (2008), pp. 153–160. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/ijc.23489.

[11] Arizona Department of Health Services. *Cancer in Arizona: Cancer Incidence and Mortality 2005-2007*. Tech. rep. Pheonix: Bureau of Public Health Statistics, 2009, p. 75.

[12] B. F. Arnold et al. "Acute Gastroenteritis and Recreational Water: Highest Burden Among Young US Children". In: *American journal of public health* (2016). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4981791/.

[13] K Asomaning et al. “Second hand smoke, age of exposure and lung cancer risk”. In: *Lung Cancer* 61.1 (2008), pp. 13–20.

[14] M. G. Baker, E. Sneyd, and N. A. Wilson. “Is the major increase in notified campylobacteriosis in New Zealand real?” In: *Epidemiology and Infection* 135 (2007), pp. 163–170. DOI: 10.1017/S0950268806006583.

[15] E. Bakker and S. Hilt. “Impact of water-level fluctuations on cyanobacterial blooms: options for management”. In: *Aquatic Ecology* 50.3 (2016).

[16] S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarichical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman and Hall/CRC, 2004, p. 452. ISBN: 1-58488-410-X.

[17] C Bankhead, S Kehoe, and J Austoker. “Symptoms associated with diagnosis of ovarian cancer: a systematic review”. In: *BJOG: An International Journal of Obstetrics and Gynaecology* 112.7 (2005), pp. 857–865.

[18] R Barnett and P Barnett. “Primary health care in New Zealand: Problems and policy approaches”. In: *Social Policy Journal of New Zealand* 21 (2004), pp. 49–66.

[19] S Becken. “Water equity – Contrasting tourism water use with that of the local community”. In: *Water Resources and Industry* 7 (2014), pp. 9–22.

[20] E. Beghi et al. “The epidemiology of ALS and the role of population-based registries. - Semantic Scholar”. In: *Biochimica et Biophysica Acta* 1762.11-12 (1970), pp. 1150–1157. URL: https://www.semanticscholar.org/paper/The-epidemiology-of-ALS-and-the-role-of-registries.-Beghi-Logroscino/c8be370cf06c31a4f992197effc2629f74f623c5.

[21] B. Bernal, C. J. Anderson, and W. J. Mitsch. "Nitrogen dynamics in two created riparian wetlands over space and time". In: *Journal of Hydrologic Engineering* 22.1 (2017).

[22] M. Bernardi. "Linkages between FAO agroclimatic data resources and the development of GIS models for control of vector-borne diseases". In: *Acta Tropica* 79.21–34 (2001).

[23] J. Besag, J. York, and A. Mollie. "Bayesian image restoration with two applications in spatial statistics". In: *Ann. Inst. Statist. Math* 43.1 (1991), pp. 1–59. URL: http://download.springer.com.ezproxy.canterbury.ac.nz/static/pdf/264/art:10.1007/BF00116466.pdf?originUrl=http://link.springer.com/article/10.1007/BF00116466\&token2=exp=1496802331~acl=/static/pdf/264/art\%3A10.1007\%2FBF0.

[24] M. J. Betancourt and M. Girolami. "Hamiltonian Monte Carlo for Hierarchical Models". In: (2013). DOI: 10.1201/b18502-5. arXiv: 1312.0906. URL: http://arxiv.org/abs/1312.0906.

[25] T Bewick, L Binning, and B Yandell. "A degree day model for predicting the emergence of swamp dodder in cranberry". In: *Journal of the American Society for Horticultural Science* 113.6 (1988), pp. 839–841.

[26] L. I. Boden and A. Ozonoff. "Capture-Recapture Estimates of Nonfatal Workplace Injuries and Illnesses". In: *Annals of Epidemiology* 18.6 (2008), pp. 500–506. ISSN: 10472797. DOI: 10.1016/j.annepidem.2007.11.003.

[27] L. Brabyn and R. Barnett. "Population need and geographical access to general practitioners in rural New Zealand". In: *The New Zealand Medical Journal* 117.1199 (2004).

[28] T. Branson and J. Krysan. "Feeding and Oviposition Behavior and Life Cycle Strategies of Diabrotica: An Evolutionary View with Implications for Pest Management". In: *Environmental EntomologyB* 10 (1981), 826–831.

[29] J Burch et al. "Diagnostic accuracy of faecal occult blood tests used in screening for colorectal cancer: a systematic review". In: *Journal of Medical Screening* 2007.14 (2007), pp. 132–137.

[30] U. C. Bureau. *American Community Survey Data*. 2019. URL: https://www.census.gov/programs-surveys/acs/data.html.

[31] P Cate. "Spread and population development of western corn rootworm (Diabrotica virgifera virgifera) in Austria." In: *Symposium on the Introduction and the Spread of Invasive Species, Humboldt University, Berlin, Germany* (2005).

[32] F. Colchero et al. "The emergence of longevous populations". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.48 (2016). URL: https://www.pnas.org/content/113/48/E7681/.

[33] J Corner et al. "Is late diagnosis of lung cancer inevitable? Interview study of patients' recollections of symptoms before diagnosis". In: *Thorax* 60 (2005), pp. 314–319.

[34] "Council Directive 2000/29/EC of 8 May 2000 on protective measures against the introduction into the Community of organisms harmful to plants or plant products and against their spread within the Community". In: *Official Journal of the European Union* (2000). URL: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32000L0029.

[35] T Crowl et al. "The spread of invasive species and infectious disease as drivers of ecosystem change". In: *Frontiers in Ecology and the Environment* 6.5 (2008).

[36] J Crowther, D. Kay, and W. M. D. "Relationships between microbial water quality and environmental conditions in coastal recreational waters: the fylde coast, UK". In: *Water Research* 35.17 (2001), pp. 4029–4038.

[37] E. C. Dasenbrook and G. S. Sawicki. "Cystic fibrosis patient registries: A valuable source for clinical research". In: *Journal of Cystic Fibrosis* 17 (2018), pp. 433–400.

[38] R. Davies-Colley and J. W. Nagels. "Effects of dairying on water quality of lowland streams in Westland and Waikato". In: *Proceedings of the New Zealand Grassland Association* (2002), pp. 207–114.

[39] J Deitloff et al. "Effects of refuges on the evolution of resistance to transgenic corn by the western corn rootworm, Diabrotica virgifera virgifera LeConte". In: *Pest Management Science* 72.1 (2016), pp. 190–198.

[40] Y Devos, L Meihls, and J Kiss. "Resistance evolution to the first generation of genetically modified Diabrotica-active Bt-maize events by western corn rootworm: management and monitoring considerations". In: *Transgenic Research* 22 (2013), pp. 269–270.

[41] T. Diepgen and V. Mahler. "The epidemiology of skin cancer". In: *British Joural of Dermatology* 146.s61 (2002), pp. 1–6. URL: https://onlinelibrary.wiley.com/doi/full/10.1046/j.1365-2133.146.s61.2.x.

[42] J Diggle and B. Matern. "On Sampling Designs for the Study of Point-Event Nearest Neighbour Distributions in R^2". In: *Scandinavian Journal of Statistics* 7.2 (1980), pp. 80–84.

[43] P Diggle and J Tawn. "Model-based geostatistics". In: *Applied Statistics* 47.3 (1998), pp. 299–350.

[44] K. Dillen et al. "The western corn rootworm, a new threat to European agriculture: opportunities for biotechnology?" In: *Pest Management Science* 66.9 (2010), pp. 965–966. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/ps.1966.

[45] N Draper and I Guttman. "Estimation of the Binomial Parameter". In: *American Statistical Association and American Society for Quality* 13.3 (1971), pp. 667–673.

[46] S. Džeroski, D. Demšar, and J. Grbović. "Predicting chemical parameters of river water quality from bioindicator data". In: *Applied Intelligence* 13.1 (2000), pp. 7–17. ISSN: 0924669X. DOI: 10.1023/A:1008323212047.

[47] EC. *Survey results for the presence of Diabrotica virgifera Le Conte in the European Union in 2011*. Tech. rep.

[48] S. Edberg, R. Karlin, and M. Allen. "Escherichia coli: the best biological drinking water indicator for public health protection". In: *Journal of Applied Microbiology* (2012). URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2672.2000.tb05338.x.

[49] J Elith et al. "A statistical explanation of MaxEnt for ecologists". In: *Diversity and Distributions* (2010), pp. 1–5.

[50] Embassy of Austria. *Climate*. URL: https://www.austria.org/climate.

[51] Environment Southland. *Monitoring rriver and beaches for faecal contaminants*. 2014. URL: https://www.es.govt.nz/repository/libraries/id:26gi9ayo517q9stt81sd/hierarchy/environment/science/science-summaries/documents/monitoring_rivers_and_beaches_for_faecal_contaminants.pdf.

[52] *EpiSurv: public health surveillance*. URL: https://surv.esr.cri.nz/episurv/index.php.

[53] *Facts on Agriculture - Taste of Austria*. URL: https://www.bmnt.gv.at/english/agriculture/Productionandmarkets/Plant-production-in-Austria/Cereal-production-and-types-of-cereals-in-Austria.html.

[54] K Falkner et al. "A zero-inflated Poisson mixture model to analyse spread and abundance of the Western Corn Rootworm in Austria". In: *Agricultural Systems* (2019). URL: https://www.sciencedirect.com/science/article/pii/S0308521X18308564#bb0195.

[55] A. Fenemor. "Water governance in New Zealand – challenges and future directions". In: *New Water Policy and Practice* 3.1 (2017), pp. 9–21.

[56] T Ferns. "Under-reporting of violent incidents against nursing staff". In: *Nursing Standard* 20.40 (2006), pp. 41–45. ISSN: 0029-6570. DOI: 10.7748/ns2006.06.20.40.41.c4178.

[57] E Feusthuber et al. "Integrated modelling of efficient crop management strategies in response to economic damage potentials of the Western Corn Rootworm in Austria". In: *Agricultural Systems* 157 (2017), pp. 93–106.

[58] G. H. Fischer and E. Paterek. *Campylobacter*. 2019. URL: https://www.ncbi.nlm.nih.gov/books/NBK537033/.

[59] D. S. Francy. "Use of predictive models and rapid medthods to nowcast bacteria llevel at coastal beachers". In: *Aquatic Ecosystem Health and Management* 12.2 (2009).

[60] D. S. Francy et al. *Developing and implementing the use of predictive models for estimating water quality at Great Lakes beaches*. 2013. URL: https://pubs.er.usgs.gov/publication/sir20135166.

[61] N Garret et al. "Statistical comparison of Campylobacter jejuni subtypes from human cases and environmental sources". In: *Journal of Applied Microbiology* 103.6 (2007), pp. 2113–2121.

[62] J Gasperino. "Gender is a risk factor for lung cancer". In: *Medical Hypotheses* 76.3 (2011), pp. 328–331.

[63] C. Gehlke and K Biehl. "Certain effects of grouping upon the size of the correlation coefficient in census tract material". In: *Journal of the American Statistical Association* 29.185 (1934), pp. 169–170.

[64] A Gelman and D. B. Rubin. "Inference from interative simulation using multiple sequences". In: *Statistical Science* 7.4 (1992), pp. 457–511.

[65] A Gelman et al. *Bayesian Data Analysis. Second edition*. Chapman and Hall/CRC, 2004.

[66] S Geng and C Jung. "Temperature-dependent development of overwinteover pupae of *Phyllonorycter ringoniella* and its spring emergence model". In: *Journal of Asia-Pacific Entomology* 21.3 (2018), pp. 829–835.

[67] Z. für Meteorologie und Geodynamik — ZAMG. *INCA: Integrated Nowcasting through Comprehensive Analysis. INCA Integrated Nowcasting Comprehensive Analysis*. 2018. URL: http://www.zamg.ac.at/cms/de/forschung/wetter/inca.

[68] I. A. Gillespie et al. "Demographic determinants for Campylobacter infection in England and Wales : implications for future epidemiological studies". In: *Epidemiology and Infection* 136.12 (2008), pp. 1717–1725. DOI: 10.1017/S0950268808000319.

[69] B. J. Gilpin et al. "The transmission of thermotolerant Campylobacter spp. to people living or working on dairy farms in New Zealand". In: *Zoonoses and Public Health* 55.7 (2008), pp. 352–360. ISSN: 18631959. DOI: 10.1111/j.1863-2378.2008.01142.x.

[70] B. Gilpin et al. *Comparison of Campylobacter jejuni genotypes from dairy cattle and human sources from the Matamata-Piako District of New Zealand*. 2008. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2672.2008.03863.x.

[71] R. E. Gliklich. *Registry Design*. 2014. URL: https://www.ncbi.nlm.nih.gov/books/NBK208632/.

[72] B Gompertz. "On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies". In: *Royal Society* 115 (1825).

[73] M Gray et al. "Adaptation and invasiveness of western corn rootworm: intensifying research on a worsening pest". In: *Annual Review of Entomology* 54 (2009), pp. 303–321.

[74] S. Greenland. "Bayesian perspectives for epidemiological research: I. Foundations and basic methods". In: *International Journal of Epidemiology* 35.3 (2006), pp. 765–775.

[75] S. Greenland and J. Robins. "Invited Commentary: Ecologic Studies - Biases, Misconceptions, and Counterexamples". In: *American Journal of Epidemiology* 139.8 (1994), pp. 747–760.

[76] A. D. Gronewold and R. L. Wolpert. "Modeling the relationship between most probable number (MPN) and colony-forming unit (CFU) estimates of fecal coliform concentration". In: *Water research* 42 (2008), pp. 3327–3334. DOI: 10.1016/j.watres.2008.04.011.

[77] K Gurney, R Cartwright, and E Gilman. "Descriptive epidemiology of gastrointestinal non-Hodgkin's lymphoma in a population-based registry". In: *British Journal of Cancer* 79.11-12 (1999), pp. 1929–1934.

[78] P. Gustafson. "On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables". In: *Statistical Science* 20.2 (2005), pp. 111–140. DOI: 10.1214/088342305000000098.

[79] C. Gutierrez and D. Kirk. "Silence speaks: The relationship between immigration and the underreporting in crime". In: *Crime and Delinquency* 63.8 (2017), pp. 926–950.

[80] G Hall et al. "Estimating foodborne gastroenteritis, Australia". In: *Emerging Infectious Diseases* 11.8 (2005), pp. 1257–1264. DOI: 10.3201/eid1108.041367. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3320479/.

[81] T Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: Data mining, inference and prediction*. Two. Springer Series in Statistics, 2008.

[82] T. Hastie and W. Fithian. "Inference from presence-only data ; the ongoing controversy". In: *Ecography* (2013), pp. 864–867. DOI: 10.1111/j.1600-0587.2013.00321.x.

[83] G. Hay. "Estimating the prevalence of drug misuse in Dundee, Scotland: An application of capture-recapture methods". In: *Journal of Epidemiology and Community Health* 50.4 (1996), pp. 469–472. ISSN: 0143005X. DOI: 10.1136/jech.50.4.469.

[84] P. D. of Health. *Pennsylvania Department of Health programs, services and health information*. URL: https://www.health.pa.gov/Pages/default.aspx.

[85] Health Navigator NZ. *Endometriosis*. 2018. URL: https://www.healthnavigator.org.nz/health-a-z/e/endometriosis/.

[86] L Hemerik, C Busstra, and P Mols. "Predicting the temperature-dependent natural population expansion of the western corn rootworm, Diabrotica virgifera". In: *Entomologia Experimentalis et Appllicata* (2004).

[87] C.-C. W. Hui-Jen Tsai and J. S. Chang. "The Epidemiology of Neuroendocrine Tumors in Taiwan: A Nation-Wide Cancer Registry-Based Study". In: *PLOS ONE* (). URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0062487.

[88] Institute of Environmental Science and Research Limited. *NOTIFIABLE AND OTHER DISEASES IN NEW ZEALAND 2008 ANNUAL SURVEILLANCE REPORT*. Tech. rep. Porir: Institute of Environmental Science and Research Limited, 2009, p. 67. URL: https://surv.esr.cri.nz/PDF_surveillance/AnnualRpt/AnnualSurv/2008AnnualSurvRpt.pdf.

[89] T. T. T. D. of Internal Affairs. *Department of Internal Affairs*. URL: https://www.dia.govt.nz/Government-Inquiry-into-Havelock-North-Drinking-Water-Report---Part-1---Overview.

[90] J. L. Jasinski. "The Effect of Victim-Offender Relationship on Reporting Crimes of Violence against Women". In: *Canadian Journal of Criminology* 16.7 (2001), pp. 393–429.

[91] J Jones. "Monitoring species abundance and distribution at the landscape scale". In: *Journal of Applied Ecology* 48.1 (2011), pp. 9–13.

[92] A. Y. Kim and J. Wakefield. *SpatialEpi: Methods and Data for Spatial Epidemiology*. R package version 1.2.2. 2016. URL: https://CRAN.R-project.org/package=SpatialEpi.

[93] J. Kiss et al. "Monitoring of western corn rootworm (Diabrotica virgifera virgifera LeConte) in Europe 1992–2003". In: *Western Corn Rootworm: Ecology and Management* (2005), pp. 29–40.

[94] A Knutson and M Muegge. "A degree-day model initiated by pheromone trap captures for managing pecan nut casebecase (Lepidoptera: Pyralidae) in pecans". In: *Horticultural Entomology* 103.3 (2010), pp. 735–743.

[95] F. I. Korennoy et al. "Spatio-temporal modeling of the African swine fever epidemic in the Russian Federation, 2007-2012". In: *Spatial and Spatio-temporal Epidemiology* 11 (2014), pp. 135–141. ISSN: 18775853. DOI: 10.1016/j.sste.2014.04.002. arXiv: 15334406. URL: http://dx.doi.org/10.1016/j.sste.2014.04.002.

[96] U. Kuhlmann and W. A. C. M. van der Burgt. "Possibilities for biological control of the western corn rootworm, Diabrotica virgifera virgifera LeConte, in Central Europe". In: *Biocontrol News and Information* 19.2 (1998), pp. 59–68.

[97] A. Lal et al. "Environmental change and enteric zoonoses in New Zealand: A systematic review of the evidence". In: *Australian and New Zealand Journal of Public Health* 39.1 (2015), pp. 63–68. ISSN: 17536405. DOI: 10.1111/1753-6405.12274.

[98] S. T. Larned et al. "Water quality in New Zealand rriver: current state and trends". In: *New Zealand Journal od Marine and Freshwater Research* 50.3 (2016), pp. 389–417.

[99] A. B. Lawson. *Bayesian disease mapping. Hierarchical modeling in spatial epidemioology. Second Edition*. CRC Press, Taylor and Francis Group, 2013.

[100] A. B. Lawson and C. Rotejanaprasert. "Childhood brain cancer in Florida: a Bayesian clustering approach". In: *Statistics and Public Policy* October 2015 (2014), p. 00. ISSN: 2330-443X. DOI: 10.1080/2330443X.2014.970247. URL: http://www.tandfonline.com/doi/abs/10.1080/2330443X.2014.970247.

[101] J. K. Lee et al. "Accuracy of fecal immunochemical tests for colorectal cancer: Systematic review and meta-analysis". In: *Annals of Internal Medicine* 160 (2014).

[102] S. Levesque et al. "Campylobacteriosis in urban versus rural areas: A case-case study integrated with molecular typing to validate risk factors and to attribute sources of infection". In: *PLoS ONE* 8.12 (2013), pp. 17–20. ISSN: 19326203. DOI: 10.1371/journal.pone.0083731.

[103] P. L. Lo and J. T. S Walker. "Annual and regional variability in adult *Dasineura mali* (apple leafcurling midge) emergence in New Zealand". In: *Horticultural Insects* 70 (2017), pp. 131–136.

[104] K. Lokar, T. Zagar, and V. Zadnik. "Estimation of the Ecological Fallacy in the Geographical Analysis of the Association of Socio-Economic Deprivation and Cancer Incidence". In: *International Journal of Environmental Research and Public Health* 16 (2019), p. 296.

[105] D. Lunn et al. "WinBUGS a Bayesian modelling framework: concepts, structure, and extensibility." In: *Journal of Statistical Software* 10 (2000), pp. 325–337. URL: https://www.mrc-bsu.cam.ac.uk/software/bugs/.

[106] Z. Macdonald. "Revisiting the dark figure: A microeconometric analysis of the under-reporting of property crime and its implications". In: *British Journal of*

Criminology 41.1 (2001), pp. 127–149. ISSN: 00070955. DOI: 10.1093/bjc/41.1.127.

[107] S Macfayden, G McDonald, and M Hill. "From species distributions to climate change adaptation: Knowledge gaps in managing invertebrate pests in broad-acre grain crops". In: *Agriculture, Ecosystems and Environment* 253 (2017), pp. 208–219.

[108] B Madon et al. "A new method for estimating animal abundance with two sources of data in capture–recapture studies". In: *Methods in Ecology and Evolution* 2.4 (2011), pp. 390–400.

[109] A. Mahr et al. "Prevalences of polyarteritis nodosa, microscopic polyangiitis, Wegener's granulomatosis, and Churg-Strauss syndrome in a French urban multiethnic population in 2000: A capture-recapture estimate". In: *Arthritis Care & Research* 51.1 (2004), pp. 92–99. ISSN: 00043591. DOI: 10.1002/art.20077. URL: http://doi.wiley.com/10.1002/art.20077.

[110] J. Maret-Ouda et al. "Nordic registry-based cohort studies: Possibilities and pitfalls when combining Nordic registry data". In: *Scandinavian Journal of Public Health* 45.17 (2017), pp. 14–19. URL: https://journals.sagepub.com/doi/full/10.1177/1403494817702336.

[111] Maryland Department of Health. *2017 Cancer Data: Cigarette Restitution Fund Program; Cancer Prevention, Education, Screening and Treatment Program.* Tech. rep. August. Baltimore: Maryland Department of Health, 2017, p. 182.

[112] L. J. Meinke et al. "Western corn rootworm (Diabrotica virgifera virgifera LeConte) population dynamics". In: *Agricultural and Forest Entomology* 11.1 (2009), pp. 29–46. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1461-9563.2008.00419.x.

[113] N Miller et al. "Multiple transatlantic introductions of the western corn rootworm". In: *Sciencemag* 310.5750 (2005).

[114] I Milnes A ad Stewart, F Clifton-Hadley, and R Davies. "Intestinal carriage of verocytotoxigenic Escherichia coli O157, Salmonella, thermophilic Campylobacter and Yersinia enterocolitica, in cattle, sheep and pigs at slaughter in Great Britain during 2003". In: *Epidemiology and Infection* 136.6 (2007).

[115] Ministry for the Environment. *Microbiological Water Quality Guidelines for Marine and Freshwater Recreational Areas*. 2002, p. 159. ISBN: 0478240910.

[116] Ministry for the Environment. *National Policy Statement for Freshwater Management 2014*. Tech. rep. New Zealand government, 2014, p. 34. URL: http://www.mfe.govt.nz/publications/fresh-water/national-policy-statement-freshwater-management-2014.

[117] Ministry of Health. *Communicable Disease Control Manual*. Tech. rep. Wellington: Ministry of Health, 2012, p. 322. URL: https://www.health.govt.nz/our-work/diseases-and-conditions/communicable-disease-control-manual/cryptosporidiosishttps://www.health.govt.nz/system/files/documents/publications/communicable-disease-control-manual-oct18.pdf.

[118] Ministry of Health. *Notifiable diseases*. 2017. URL: http://www.health.govt.nz/our-work/diseases-and-conditions/notifiable-diseases (visited on 06/07/2017).

[119] P. Mischler et al. "Environmental and socio-economic risk modelling for Chagas disease in Bolivia". In: *Geospatial Health* 6.3 SUPPL. (2012). ISSN: 18271987. DOI: 10.4081/gh.2012.123.

[120] R. M. Monaghan et al. "Linkages between land management activities and water quality in an intensively farmed catchment in southern New Zealand". In: *Agricultural Ecosystems and Environment* (2007).

[121] D. Moore et al. *The Economic Costs of the Havelock North August 2016 Waterborne Disease Outbreak*. Tech. rep. August 2016. Sapere research group, 2017, p. 56. URL: https://www.health.govt.nz/system/files/documents/publications/havelock_north_outbreak_costing_final_report_-_august_2017.pdf.

[122] E. Moreno and F. J. Girón. "Estimating with incomplete count data A Bayesian approach". In: *Journal of Statistical Planning and Inference* 66.1 (1998), pp. 147–159. ISSN: <null>. URL: papers2://publication/uuid/7490F2C5-8DEA-4014-BB6A-5EDAA263F90B.

[123] P. Mullner et al. "Assigning the source of human campylobacteriosis in New Zealand: A comparative genetic and epidemiological approach". In: *Infection, Genetics and Evolution* 9.6 (2009), pp. 1311–1319. ISSN: 15671348. DOI: 10.1016/j.meegid.2009.09.003. URL: http://www.sciencedirect.com/science/article/pii/S1567134809001981.

[124] B Nambam, S Aggarwal, and A Jain. "Latent autoimmune diabetes in adults: A distinct but heterogeneous clinical entity". In: *World Journal of Diabetes* 1.4 (2010), pp. 111–115.

[125] New Zealand Law Resources. *Health Act 1956*. 1956.

[126] G. L. Nichols et al. "Campylobacter epidemiology : a descriptive study reviewing 1 million cases in England and Wales between 1989 and 2011". In: *BMJ Open* 2 (2012), pp. 1–13. DOI: 10.1136/bmjopen-2012-001179.

[127] J Nichols and B Williams. "Monitoring for conservation". In: *Trends in Ecology and Evolution* 21.12 (2006), pp. 668–673.

[128] G. L. D. Oliveira, R. H. Loschi, and R. M. Assunção. "A random-censoring Poisson model for underreported data". In: *Statistics in Medicine* 36.August (2017), pp. 4873–4892. DOI: 10.1002/sim.7456.

[129] U. Oliveira et al. "The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity". In: *Diversity and Distributions* 22 (2016), pp. 1232–1244. URL: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ddi.12489.

[130] B Oporto et al. "Prevalence and strain diversity of thermophilic campylobacters in cattle, sheep and swine farms". In: *Journal of Applied Microbiology* 104.4 (2007), pp. 977–984.

[131] T. B. Parr et al. "Potential roles of past, present, and future urbanization characteristics in producing varied stream responses". In: *Freshwater Science* 35.1 (2016).

[132] I. Pattis et al. *Annual Report Concerning Foodborne Disease in New Zealand 2016, 2017: ESR Client Report FW17008, Christchurch, New Zealand.* Tech. rep. May. 2017, p. 147.

[133] L Payne et al. "'Did you have flu last week?' A telephone survey to estimate point prevalence of influenza in the Swedish population". In: *Euro surveillance* 10.12 (2005), pp. 5–6. DOI: 10.2807/esm.10.12.00585-en.

[134] J. L. Pearce and M. S. Boyce. "Modelling distribution and abundance with presence-only data". In: *Journal of Applied Ecology* 43 (2006), pp. 405–412. DOI: 10.1111/j.1365-2664.2005.01112.x.

[135] E. J. Pebesma and R. S. Bivand. "Classes and methods for spatial data in R". In: *R News* 5.2 (2005), pp. 9–13. URL: https://CRAN.R-project.org/doc/Rnews/.

[136] F. Pezzella, M. Fetzer, and T. Keller. "The dark figure of hate crime underreporting". In: *American Behavioral Scientist* (2019).

[137] K. Phillips et al. "Combining Watchman left atrial appendage closure and catheter ablation for atrial fibrillation: multicentre registry results of feasibility and safety during implant and 30 days follow-up". In: *European Society of Cardiology* 20 (2018). URL: https://academic.oup.com/europace/article/20/6/949/3920546/.

[138] S. J. Phillips, R. P. Anderson, and R. E. Schapire. "Maximum entropy modeling of species geographic distributions." In: *Ecological Modelling* (2006).

[139] S. B. Phillips et al. "Modelling and analysis of the atmospheric nitrogen deposition in North Carolina". In: *International Journal of Global Environmental Issues* 6.2-3 (2006), pp. 231–252. ISSN: 14666650. DOI: 10.1016/j.ecolmodel.2005.03.026. arXiv: 11265.

[140] S. Phillips and M Dudik. "Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation". In: *Ecography* 31 (2008), pp. 161–175.

[141] P Pinsky. "Racial and ethnic differences in lung cancer incidence: how much is explained by differences in smoking patterns? (United States)". In: *Cancer Causes and Control* 17.9 (2006), pp. 1017–1024.

[142] A. Prüss. "Review of epidemiological studies on health effects from exposure to recreational water". In: *International Journal of Epidemiology* 27.1 (1998), pp. 1–9. ISSN: 0300-5771. DOI: 10.1093/ije/27.1.1. URL: http://www.ncbi.nlm.nih.gov/pubmed/9563686.

[143] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2015. URL: http://www.R-project.org/.

[144] E. Radzikowska, K. Roszkowski, and K. Głaz. "Lung cancer in women: age, smoking, histology, performance status, stage, initial treatment and survival. Population-based study of 20,561 cases". In: *Annals of Oncology* 13.7 (2002), pp. 1087–1093.

[145] B. Y. A. E. Raftery. "Inference for the binomial N parameter: A hierarchical Bayes approach". In: *Biometrika* (1988), pp. 223–228.

[146] J. Ranta et al. "Bayesian risk assessment for Salmonella in egg laying flocks under zero apparent prevalence and dynamic test sensitivity". In: *Journal de la Societe Francaise de Statistique* 154.3 (2013).

[147] J. a. Razzak and S. P. Luby. "Estimating deaths and injuries due to road traffic accidents in Karachi, Pakistan, through the capture-recapture method". In: *International Journal of Epidemiology* 27.5 (1998), pp. 866–870. ISSN: 03005771. DOI: 10.1093/ije/27.5.866.

[148] C. Reed et al. "Estimates of the prevalence of pandemic (H1N1) 2009, United States, april-july 2009". In: *Emerging Infectious Diseases* 15.12 (2009), pp. 2004–2007. ISSN: 10806040. DOI: 10.3201/eid1512.091413.

[149] "Repealing Decision 2003/766/EC on emergency measures to prevent the spread within the Community of Diabrotica virgifera Le Conte". In: *Official Journal of the European Union* (2014).

[150] S. Richardson et al. "Interpreting Posterior Relative Risk Estimates in Disease-Mapping Studies". In: *Environmental Health Perspectives* 112.9 (2003), pp. 1016–1025.

[151] E. Rind and J. Pearce. "The spatial distribution of campylobacteriosis in New Zealand, 1997-2005". In: *Epidemiology and Infection* 138 (2010), pp. 1359–1371. DOI: 10.1017/S095026881000018X.

[152] B Robertson et al. "BAS: Balanced accepted sampling of natural resources". In: *Biometrics* 69.3 (2013), pp. 776–784.

[153] J. A. Royle. "N-mixture models for estimating population size from spatially replicated counts". In: *Biometrics* 60.1 (2004), pp. 108–115.

[154] J. A. Royle, J. D. Nicols, and M Kerry. "Modelling occurrence and abundance of species when detection is imperfect". In: *Oikos* 110.2 (2005), pp. 353–359.

[155] J. A. Royle and J. D. Nichols. "ESTIMATING ABUNDANCE FROM REPEATED PRESENCE â€" ABSENCE data or point counts". In: *Ecology* 84.3 (2003), pp. 777–790. ISSN: 0012-9658. DOI: 10.1890/0012-9658(2003)084[0777:EAFRPA]2.0.CO;2.

[156] J. A. Royle et al. "Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions". In: *Methods in Ecology and Evolution* 3 (2012), pp. 545–554. DOI: 10.1111/j.2041-210X.2011.00182.x.

[157] D. B. Rubin. "Inference and missing data". In: *Biometrika* 63.3 (1976), pp. 581–592.

[158] G. Salmon. "Sustainability issues in New Zealand agriculture – and possibilities for collaborative resolution of them". In: *Proceedings of the New Zealand Grassland Association* (2007), pp. 11–15.

[159] J Santosh and P. Crampton. "Primary health care in New Zealand: Who has access?" In: *Health Policy* 93.1 (2009), pp. 1–10.

[160] E Scallan et al. “Hospitalisations due to bacterial gastroenteritis: A comparison of surveillance and hospital discharge data”. In: *Epidemiology and Infection* 146 (2018), pp. 954–960.

[161] C. P. Schmertmann and M. R. Gonzaga. “Bayesian Estimation of Age-Specific Mortality and Life Expectancy for Small Areas With Defective Vital Records”. In: *Demography* 55.4 (2018), pp. 1363–1388. ISSN: 15337790. DOI: 10.1007/s13524-018-0695-2.

[162] A. Sears et al. “Marked Campylobacteriosis Decline after Interventions Aimed at Poultry, New Zealand”. In: *Emerging Infectious Diseases* 17.6 (2011), p. 18. DOI: 10.3201/eid1706.101272. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3358198/pdf/10-1272_finalR.pdf.

[163] F Sinabell et al. *Maisanbau in Österreich: Ökonomische Bedeutung und pflanzenbauliche Herausforderungen 97*. URL: https://ideas.repec.org/b/wfo/wstudy/58147.html.

[164] M. T Smith et al. “Dispersal and spatiotemporal dynamics of asian longhorned beetle (Coleoptera: Cerambycidae) in China”. In: *Environmental Entomology* 33.2 (2004), pp. 435–442.

[165] A. Soleimani et al. “Spatial analysis of common gastrointestinal tract cancers in counties of Iran.” In: *Asian Pacific journal of cancer prevention : APJCP* 16.9 (2015), pp. 4025–4029. ISSN: 1513-7368 (Print). DOI: 10.7314/APJCP.2015.16.9.4025.

[166] J. Spencer et al. “Behaviour and ecology of the western corn rootworm (Diabrotica virgifera virgifera LeConte)”. In: *Agriculture and Forest Entomology* 11 (2009), 9–27.

[167] S. E. F. Spencer et al. "The spatial and temporal determinants of campylobacteriosis notifications in New Zealand , 2001-2007". In: *Epidemiology and Infection* 140 (2011), pp. 1663–1677. DOI: 10.1017/S0950268811002159.

[168] S. E. F. Spencer et al. "Spatial and Spatio-temporal Epidemiology The detection of spatially localised outbreaks in campylobacteriosis notification data". In: *Spatial and Spatio-temporal Epidemiology* 2.3 (2011), pp. 173–183. ISSN: 1877-5845. DOI: 10.1016/j.sste.2011.07.008. URL: http://dx.doi.org/10.1016/j.sste.2011.07.008.

[169] D. J. Spiegelhalter et al. "Bayesian measures of model complexity and fit". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 64.4 (2002), pp. 583–616. ISSN: 13697412. DOI: 10.1111/1467-9868.00353. arXiv: arXiv:1011.1669v3.

[170] K Stanley and K Jones. "Catlle and sheep farms as reservoirs of *Campylobacter*". In: *Journal of Applied Microbiology* 94.1 (2003), pp. 104–114.

[171] Statistics New Zealand. *Classifications and related statistical standards*. URL: http://www.stats.govt.nz/methods/classifications-and-standards/classification-related-stats-standards.aspx (visited on 06/07/2017).

[172] Statistics New Zealand. *Population mobility of urban / rural profile areas*. 2006. URL: http://www.stats.govt.nz/browse_for_stats/population/Migration/internal-migration/mobility-urban-rural-areas.aspx (visited on 06/07/2017).

[173] Statistics New Zealand. *Urban/Rural Profile (experimental) Classification Categories*. 2006. URL: http://www.stats.govt.nz/methods/classifications-and-standards/urban-rural-profile-experimental-class-categories.aspx (visited on 07/04/2017).

[174] D Stevenson et al. "Physiological time model for predicting adult emergence of western corn rootworm (CFU)oleoptera: chrysomelidae) in the Texas high plains". In: *Journal of Economic Entomology* 101.5 (2008), p. 1584.

[175] R. T. Stidson, C. A. Gray, and C. D. McPhail. "Development and use of modelling techniques for real-time bathing water quality predictions". In: *Water and Environment Journal* 26.1 (Mar. 2012), pp. 7–18. ISSN: 17476585. DOI: 10.1111/j.1747-6593.2011.00258.x. URL: http://doi.wiley.com/10.1111/j.1747-6593.2011.00258.x.

[176] M Stone et al. "Incorrect and incomplete coding and classification of diabetes: a systematic review". In: *Diabetic Medicine* 27 (2010), pp. 491–497.

[177] O. Stoner et al. "A hierarchical framework for correcting underreporting in count data". In: *Journal of the American Statistical Association* 0 (2019), pp. 1–12.

[178] M. Stroh et al. "Are female offenders underreported compared to male offenders? A German-Greek comparison of crime reporting, rating of offence seriousness and personal experiences of victimisation". In: *European Journal on Criminal Policy and Research* 22.4 (2016), pp. 635–653.

[179] S. Sturtz, U. Ligges, and A. Gelman. "R2WinBUGS: A Package for Running WinBUGS from R". In: *Journal of Statistical Software* 12.3 (2005), pp. 1–16. URL: http://www.jstatsoft.org.

[180] M Szalai et al. "Simulating crop rotation strategies with a spatiotemporal lattice model to improve legislation for the management of the maize pest Diabrotica virgifera virgifera". In: *Agricultural Systems* 124 (2014), pp. 39–50.

[181] E Tacconelli et al. "Epidemiology, medical outcomes and costs of catheter-related bloodstream infections in intensive care units of four European countries:

literature- and registry-based estimates". In: *Journal od Hospital Infection* 72.2 (2009), pp. 97–103.

[182] S Tanaskovic et al. "Influence of artifical infestation with western corn rootwork eggs on maize morpholohy". In: *Savetovananje o Biotehnologij* (2018), pp. 377–383.

[183] W Thoe and H. W. Lee. "Daily forecasting of Hong Kong beach water quality using multiple linear regression models". In: *Journal of Environmental Engineering* 140.2 (2014).

[184] W Thoe et al. "Predicting water quality at Santa Monica Beach: Evaluation of five different models for public notification of unsafe swimming conditions." In: *Water research* 67C (2014), pp. 105–117. ISSN: 1879-2448. DOI: 10.1016/j.watres.2014.09.001. URL: http://www.ncbi.nlm.nih.gov/pubmed/25262555.

[185] N Tinsley et al. "Estimation of efficacy functions for products used to manage corn rootworm larval injury". In: *Journal of Applied Entomology* 140.6 (2016), pp. 414–425.

[186] K. M. C. Tjorve and E. Tjorve. "The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the Unified-Richards family". In: *PLOS ONE* (2017).

[187] S. Toepfer and U. Kuhlmann. "Natural mortality factors acting on western corn rootworm populations: a comparison between the United States and Central Europe, in: Vidal, S., Kuhlmann, U., Edwards, C.R. (Eds.), Western Corn Rootworm: . CABI Pub, Wallingford, Oxfordshire, UK, pp. 95–120." In: *2005* 95-120 (Ecology and Management).

[188] S. Vidal, U. Kuhlmann, and C. Edwards. *Western Corn Rootworm: Ecology and Management*. CABI Pub., 2004. ISBN: 9780851990705. URL: https://books.google.co.nz/books?id=QJSji7Q5OyUC.

[189] T. J. Wade et al. "Associated Gastrointestinal Illness Published by : Rapidly Measured Indicators of Recreational Water Quality Are Predictive of Swimming-Associated Gastrointestinal Illness". In: *The National Institute of Health Sciences* 114.1 (2006), pp. 24–28. DOI: 10.1289/ehp.8273.

[190] J. Wakefield. "Disease mapping and spatial regression with count data". In: *Biostatistics* 8.2 (2007), pp. 158–183.

[191] T. Waldhoer, M. Wald, and H. Heinzl. "Analysis of the spatial distribution of infant mortality by cause of death in Austria in 1984 to 2006." In: *International journal of health geographics* 7 (2008), p. 21. ISSN: 1476-072X. DOI: 10.1186/1476-072X-7-21.

[192] F. Wang et al. "Accommodating the ecological fallacy in disease mapping in the absence of individual exposures". In: *Statistics in Medicine* August (2017). DOI: 10.1002/sim.7494.

[193] D. Warton. *Some big news about MAXENT*. 2013. URL: https://methodsblog.com/2013/02/20/some-big-news-about-maxent/.

[194] R. Webster and M. Oliver. *Geostatistics for environmental scientists*. John Wiley & Sons, Ltd, 2001.

[195] *What's the Difference between MPN and CFU?: The IDEXX Water Testing Solutions Blog*. URL: https://www.idexx.com/en/blogs/idexx-water-testing-solutions/what-s-the-difference-between-mpn-and-cfu/.

[196] C. K. Wikle, A Zammit-Mangion, and N. Cressie. *Spatio-temporal statistics with R*. Ed. by J Chambers et al. Chapman and Hall/CRC, 2019.

[197] R Winkelmann. "Markov Chain Monte Carlo analysis of underreported count data with an application to worker absenteeism". In: *Empirical Economics* 21.4 (1996), pp. 575–587. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0346976438\&partnerID=40\&md5=32f4e0a7edccb664b01e16b66844c0b6.

[198] H WInston, C. M. Harvey, and C. F Harvey. "Arsenic in groundwater in Bangladesh: A geostatistical and epidemiological framework doe evaluating health effects and potential remedies". In: *Water Resources Research* 39.6 (2003).

[199] C. Yackulic et al. "Presence-only modelling using MAXENT: when can we trust the inferences?" In: *Methods in Ecology and Evolution* 4.3 (2012), pp. 236–243.

[200] S Yang, J Logan, and D Coffey. "Mathematical formulae doe calculating base temperature for growing degree days". In: *Agricultural and Forest Meteorology* 74 (1994), pp. 61–74.

[201] L. Zhuang and N. Cressie. "Spatio-temporal modeling of sudden infant death syndrome data". In: *Statistical Methodology* 9.1-2 (2012), pp. 117–143. ISSN: 15723127. DOI: 10.1016/j.stamet.2011.01.006. URL: http://dx.doi.org/10.1016/j.stamet.2011.01.006.